HANDBOOKS IN ECONOMICS 27

HANDBOOK OF LAW AND ECONOMICS

VOLUME 1

Editors:
A. Mitchell Polinsky
Steven Shavell

NORTH-HOLLAND

HANDBOOK OF LAW AND ECONOMICS
VOLUME 1

HANDBOOKS
IN
ECONOMICS

27

Series Editors

KENNETH J. ARROW
MICHAEL D. INTRILIGATOR

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
North-Holland is an imprint of Elsevier

N·H

HANDBOOK OF LAW AND ECONOMICS

VOLUME 1

Edited by

A. MITCHELL POLINSKY
Josephine Scott Crocker Professor of Law and Economics
Stanford University, CA, USA

and

STEVEN SHAVELL
Samuel R. Rosenthal Professor of Law and Economics
Harvard University, MA, USA

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
North-Holland is an imprint of Elsevier

N·H

North-Holland is an imprint of Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
Linacre House, Jordan Hill, Oxford OX2 8DP, UK

First edition 2007

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN-13: 978-0-444-51235-2

ISSN: 0169-7218 (Handbooks in Economics Series)
ISSN: 1574-0730 (Handbook of Law and Economics series)

For information on all North-Holland publications
visit our website at books.elsevier.com

Printed and bound in the United Kingdom

07 08 09 10 11 10 9 8 7 6 5 4 3 2 1

INTRODUCTION TO THE SERIES

The aim of the *Handbooks in Economics* series is to produce Handbooks for various branches of economics, each of which is a definitive source, reference, and teaching supplement for use by professional researchers and advanced graduate students. Each Handbook provides self-contained surveys of the current state of a branch of economics in the form of chapters prepared by leading specialists on various aspects of this branch of economics. These surveys summarize not only received results but also newer developments, from recent journal articles and discussion papers. Some original material is also included, but the main goal is to provide comprehensive and accessible surveys. The Handbooks are intended to provide not only useful reference volumes for professional collections but also possible supplementary readings for advanced courses for graduate students in economics.

KENNETH J. ARROW and MICHAEL D. INTRILIGATOR

This page intentionally left blank

CONTENTS OF THE HANDBOOK

VOLUME 1

PART I: PRIMARY ELEMENTS OF THE LEGAL SYSTEM

PART II: ADDITIONAL AREAS OF THE LEGAL SYSTEM

VOLUME 2

PART II: ADDITIONAL AREAS OF THE LEGAL SYSTEM—continued

This page intentionally left blank

PREFACE

Law can be viewed as a body of rules and legal sanctions that channel behavior in socially desirable directions—for example, by encouraging individuals to take proper precautions to prevent accidents or by discouraging competitors from colluding to raise prices. The incentives created by the legal system are thus a natural subject of study by economists. Moreover, given the importance of law to the welfare of societies, the economic analysis of law merits prominent treatment as a subdiscipline of economics. Our hope is that this two volume Handbook will foster the study of the legal system by economists.

The origins of law and economics may be traced to eighteenth century writings on crime by Beccaria (1767) and Bentham (1789). The modern incarnation of the field dates from the 1960s: Coase (1960) on property rights, externalities, and bargaining; Calabresi (1961, 1970) on liability rules and accident law; Demsetz (1967) on the emergence of property rights; and Becker (1968) on crime. Of great significance was Posner (1972), the first application of economic analysis to the body of law as a whole (Posner also authored numerous influential articles on specific legal topics).

This early writing in law and economics was mainly informal and emphasized basic subject areas of law. Later scholarship began to include formal work, notably by Brown (1973), Diamond (1974), Spence (1977), and Shavell (1980a) on liability rules and accidents, Polinsky (1979) on property rights and liability rules, Barton (1972) and Shavell (1980b) on contract law, Landes (1971) and Gould (1973) on litigation behavior, and, following Becker (1968), Polinsky and Shavell (1979) on law enforcement. Law and economics scholarship also expanded into other subject areas, with corporate law receiving the most attention—see, for example, early contributions by Bebchuk (1985), Easterbrook and Fischel (1991) (synthesizing previously written articles), Gilson (1981), and Manne (1965). Additionally, empirical research was undertaken, initially mostly in the area of crime, and subsequently in many other fields as well, especially corporate law.

The purpose of this Handbook is to provide economists with a systematic introduction to and survey of research in the field of law and economics. The Handbook contains 22 chapters and is organized into three main parts. Part I deals with the building blocks of the legal system: property law; contract law; accident law (torts); litigation (including aspects of civil procedure); and public enforcement of law (including criminal law). Part II treats other prominent areas of law: corporate law; bankruptcy law; antitrust law; regulation (of externalities, natural monopolies, and network industries); employment and labor law; antidiscrimination law; intellectual property law; environmental law; and international law. Part III addresses three additional topics: norms and the law; the experimental study of law; and political economy and the law. Most of the chapters are theoretically-oriented, but many mention relevant empirical work and three focus on empirical research (on civil law, public law enforcement, and corporate law).

The first volume of the Handbook includes all of Part I and several chapters from Part II. The second volume contains the remaining chapters of Part II and all of Part III.

We are grateful to Kenneth Arrow and Michael Intriligator for encouraging the development of the Handbook and for their substantive suggestions about it; to Valerie Teng and Mark Newson of Elsevier for their able assistance with the administrative tasks associated with its production; and to the John M. Olin Foundation, through our respective institutions' law and economics programs, for supporting our preparation of it.

A. Mitchell Polinsky & Steven Shavell

References

Barton, J.H. (1972). "The Economic Basis of Damages for Breach of Contract". Journal of Legal Studies 1, 277–304.

Bebchuk, L.A. (1985). "Toward Undistorted Choice and Equal Treatment in Corporate Takeovers". Harvard Law Review 98, 1695–1808.

Beccaria, C. (1767). "On Crimes and Punishments, and Other Writings". Cambridge University Press, New York. Editor, R. Bellamy, Translator, R. Davies et al., 1995.

Becker, G.S. (1968). "Crime and Punishment: An Economic Approach". Journal of Political Economy 76, 169–217.

Bentham, J. (1789). "An Introduction to the Principles of Morals and Legislation". In: The Utilitarians. Anchor Books, Garden City, NY, 1973.

Brown, J.P. (1973). "Toward an Economic Theory of Liability". Journal of Legal Studies 2, 323–349.

Coase, R.H. (1960). "The Problem of Social Cost". Journal of Law and Economics 3, 1–44.

Calabresi, G. (1961). "Some Thoughts on Risk Distribution and the Law of Torts". Yale Law Journal 70, 499–553.

Calabresi, G. (1970). "The Costs of Accidents: A Legal and Economic Analysis". Yale University Press, New Haven.

Demsetz, H. (1967). "Toward a Theory of Property Rights". American Economic Review: Papers and Proceedings 57, 347–359.

Diamond, P.A. (1974). "Single Activity Accidents". Journal of Legal Studies 3, 107–164.

Easterbrook, F.H., Fischel, D.R. (1991). "The Economic Structure of Corporate Law". Harvard University Press, Cambridge.

Gilson, R. (1981). "A Structural Approach to Corporations: The Case Against Defensive Tactics in Tender Offers". Stanford Law Review 33, 819–891.

Gould, J.P. (1973). "The Economics of Legal Conflicts". Journal of Legal Studies 2, 279–300.

Landes, W.M. (1971). "An Economic Analysis of the Courts". Journal of Law and Economics 14, 61–107.

Manne, H.G. (1965). "Mergers and the Market for Corporate Control". Journal of Political Economy 73, 110–120.

Polinsky, A.M. (1979). "Controlling Externalities and Protecting Entitlements: Property Right, Liability Rule, and Tax Subsidy Approaches". Journal of Legal Studies 8, 1–48.

Polinsky, A.M., Shavell, S. (1979). "The Optimal Tradeoff between the Probability and Magnitude of Fines". American Economic Review 69, 880–891.

Posner, R.A. (1972). "Economic Analysis of Law", 1st edn. Little, Brown and Company, Boston.

Shavell, S. (1980a). "Strict Liability versus Negligence". Journal of Legal Studies 9, 1–25.

Shavell, S. (1980b). "Damage Measures for Breach of Contract". Bell Journal of Economics 11, 466–490.

Spence, A.M. (1977). "Consumer Misperceptions, Product Failure and Producer Liability". Review of Economic Studies 44, 561–572.

CONTENTS OF VOLUME 1

Chapter 6

The Theory of Public Enforcement of Law

Chapter 7

Empirical Study of Criminal Punishment

Chapter 8

Environmental Law

Chapter 9

Regulation of Health, Safety, and Environmental Risks

Chapter 10

Taxation

PART I:

PRIMARY ELEMENTS OF THE LEGAL SYSTEM

This page intentionally left blank

Chapter 1

CONTRACT LAW

BENJAMIN E. HERMALIN
Walter A. Haas School of Business and Department of Economics, University of California, Berkeley

AVERY W. KATZ
School of Law, Columbia University

RICHARD CRASWELL
School of Law, Stanford University

Contents

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)01001-8

Abstract

This chapter surveys major issues arising in the economic analysis of contract law. It begins with an introductory discussion of scope and methodology, and then addresses four main topics that correspond to the major doctrinal divisions of the law of contracts. These divisions include freedom of contract (the extent of private power to create binding obligations), formation of contracts (the procedural mechanics of exchange, and the rules that govern pre-contractual behavior), contract interpretation (the consequences that follow when agreements are ambiguous or incomplete), and enforcement of contractual obligations (the choice between private and public enforcement, and the legal remedies that follow from breach of contract). In each of these sections, we provide an economic analysis of relevant legal rules and institutions, and of the connections between legal arrangements and corresponding topics in microeconomic theory, such as welfare economics and the theory of contracts.

Keywords

freedom of contract, Coase Theorem, contract formation, incomplete contracts, breach of contract, remedies for breach

JEL classification: D23, D6, D82, D86, K12, K41, L14, L51

1. Introduction

The essence of a free-market economy is the ability of private parties to enter into voluntary agreements that govern the economic exchange between them. Consequently, the law that governs such agreements is critical to the functioning of such economies. While the law of property determines the configuration of entitlements that form the basis of production and exchange, and the law of torts protects those entitlements from involuntary encroachment and expropriation, it is contract law that sets the rules for exchanging individual claims to entitlements and, thus, determines the extent to which society is able to enjoy the gains from trade. Accordingly, economists interested in the welfare properties of institutions in particular, or the micro-foundations of exchange generally, have good reason to take account of the law of contracts.

This chapter, accordingly, surveys the main issues arising in the economic analysis of contract law. We discuss both the main features of contract law as they relate to the problem of economic exchange, and how relevant legal rules and institutions can be analyzed from an economic perspective. In this introductory section, we set out the basic scope, methodology, and organization of the discussion to follow. Subsection 1.1 discusses why formal and informal contracts exist, and what economic functions they serve. Subsection 1.2 distinguishes between positive and normative issues in the economic analysis of contract law, and discusses some methodological problems associated with applying standard economic analysis to legal institutions and legal scholarship. Subsection 1.3 identifies limits on the chapter's scope and provides bibliographic recommendations for material we don't cover; and subsection 1.4 sets out the organization of the remainder of the chapter.

A caveat is in order at the outset: although it is conventional to present contract law as a discrete field, one should understand that, to a significant extent, the operation of the rules and institutions discussed below will depend on other aspects of the law, including the fields of tort, bankruptcy, procedure, and evidence. Lawyers have a cliché that describes this interdependence; they say that "the law is a seamless web." It is useful to keep in mind that many issues that economists would regard as contractual, including some important limits on contractual freedom, are governed not by contract but by tort law. Additionally, the rules relating to certain categories of exchange, such as consumer, employment, insurance, and information-licensing contracts, have developed specialized content to the point that they are often treated as distinct legal fields. Finally, the practical ability of contracting parties to assert their legal entitlements depends importantly on the procedural rules that govern courts and other enforcement institutions. Many of the specific features of contract law that we discuss below cannot be understood except as a response to the costs and other limitations of such institutions.

1.1. The economic motive for contracts

In a neoclassical exchange economy of the sort analyzed by Walras (1874) or Arrow–Debreu (Arrow and Debreu, 1954 and Debreu, 1959), there is little need for either

		Seller	
		Buy insurance	Don't buy
Buyer	Buy insurance	$-p_b, -p_s$	$-p_b, 0$
	Don't buy	$0, -p_s$	$-\ell_b, -\ell_s$

Figure 1. Coordination of insurance payments.

contracts or contract law, since buyers and sellers can exploit all gains from trade through spot transactions. Indeed, in spot markets, such as public bazaars, the parties manage reasonably well without formal contracting. Contracting becomes worthwhile when there is a temporal element to exchange or one party, at least, is unsure as to what her counterparty will do. For example, when the item to be exchanged needs to produced or the service being rendered takes time. Absent a contract, the parties could be reluctant to trust each other to complete the agreed upon exchange at the called-upon time, and thus valuable exchange is forgone. Conversely, contracts can be worthwhile even in non-exchange settings, as when advance commitment enhances the value of a gift by enabling reliance by the beneficiary (R. Posner, 1977, and Shavell, 1991) or when a supplier's commitment to remain in a market notwithstanding short-run losses deters competitive entry by rivals (*e.g.*, Rasmusen et al., 1991) or encourages entry by producers of complementary goods. The central question then becomes why commitment is valuable, to which there are several answers.

1.1.1. Coordination

The most straightforward reason to use contracts is to coordinate independent actions in situations of multiple equilibria. As an illustration, consider the game depicted in Figure 1, in which a buyer, b, and a seller, s, are independently deciding whether to purchase insurance against the loss of a good in transit. The parties' payoffs net of this decision are normalized to zero. Denote the insurance premium when purchased by party i by p_i and the *expected* loss suffered by party i when no insurance is purchased by ℓ_i. Assume that $\ell_b + \ell_s > p_b > p_s$; that is, going uninsured is more expensive than buying insurance, but it is cheaper for the seller to buy the insurance than for the buyer to do so. (This is perhaps because the seller, who packages the goods for shipment, is better able to control moral hazard and thus can obtain a better rate.)

If the parties make their choices independently, there are three Nash (1951) equilibria to this game: one in which the seller buys insurance, one in which the buyer buys insurance, and a mixed-strategy equilibrium in which both buy insurance with positive probability. Of these, the first equilibrium is the efficient one (and it is also more efficient than any disequilibrium outcome). A contract to play this efficient equilibrium could serve to ensure this outcome is achieved.

Formal contracts are of course not the only way for parties to coordinate among multiple equilibria. The efficiency of the seller-buys equilibrium could make it a focal point for the parties (Schelling, 1960). The literature on "cheap talk" (*e.g.*, Farrell, 1987a, 1993) suggests that such coordination can, in principle, also be achieved by having the parties announce their intentions in advance.

In actual institutional settings, however, contracts offer more stability than focal points or mere announcements. In particular, they provide permanent authoritative records that can be used by parties who suffer from imperfect recall or by those who need to delegate performance to their agents or successors. Note that when contracts are used for pure coordination purposes, they are self-enforcing in the sense that it is in each party's private interest *ex post* to comply with the chosen equilibrium. Hence, those designing such contracts can devote most of their attention to problems of formation and interpretation, and relatively little attention to problems of enforcement. This coordination function of contracts has been rather less discussed in the law and economics literature than the incentive mechanism functions discussed below, but it may be by far the most important purpose that contracts serve in practice. As Myerson (2004) suggests, coordination games could be the best models through which to understand legal institutions generally.

1.1.2. Implementing exchange over time

A second reason for using contracts is to implement exchanges that depend on future events. For instance, consider an insurance contract that covers a loss that occurs in state 1, but not in state 0. Under this contract, the insured pays a premium to the insurer, in exchange for a casualty payment received in state 1, but not in state 0.

In the standard model of risk allocation, goods are state-contingent commodities; for example, an apple in state 0 is considered a different good than an apple in state 1. Treating goods as state-contingent commodities has the advantage of allowing direct application of standard analyses of exchange, but has the disadvantage of abstracting from real institutional issues. In particular, exchange that is mutually beneficial *ex ante* may not be *mutually* beneficial *ex post*. In the insurance example, the insured will not wish to pay the premium *ex post* if state 0 is realized, and the insurer will not wish to make the casualty payment *ex post* if state 1 is realized. Such exchanges cannot, therefore, be implemented in spot markets and require some form of advance commitment.

In the typical insurance context, the transaction is motivated by the insured's risk aversion. But the need to contract across different states is more general than that and is not reliant on risk aversion. For instance, consider purely speculative exchange between risk-neutral parties, in which trade is motivated by differences of opinion regarding the probability of future events. If one trader thinks the price of orange juice will rise next year and another thinks it will fall, they can make themselves better off *ex ante* by entering into a forward exchange in which the second promises to deliver to the first. As with the insurance contract, this exchange requires a commitment mechanism since *ex post* one of the parties is sure to regret the deal.

An analogous problem arises with the rental of capital assets or the extension of credit. Even though the owner of an asset may not be its highest-value user, she may be unwilling to yield possession to a higher-value user for fear that she will be unable to recover it at the end of the rental period. The law of property provides a partial solution to this problem by entitling the owner to reclaim her asset, but evidentiary difficulties make this alternative an imperfect one (as illustrated by the maxim, "possession is nine tenths of the law"). The party in possession could claim, for instance, that the transaction was a gift or a sale, or that the agreed lease period had not yet expired. In such settings, a contract that specifies the parties' relative rights and duties makes the borrower's promise to return the asset more credible, facilitating exchange.

More generally, some form of commitment is necessary in any exchange in which performance is sequential, because the party who performs first is effectively extending credit to the party who performs second. It may be possible to structure the exchange so that each stage of a party's performance is timed to coincide with the performance of her counterparty (*e.g.*, an installment sale of goods in which each shipment is delivered C.O.D.), but in many instances such timing may be infeasible or costly. For instance, consider a grocery that requires regular delivery of a perishable commodity such as milk. The costs of making and receiving payment (keeping cash on hand, updating accounts, preventing embezzlement, etc.) generate substantial scale economies if disbursements for multiple shipments are combined into a single monthly payment.

Contracts can also be useful in situations of hidden information. In Akerlof's (1970) lemons model, for example, adverse selection can prevent efficient exchange when the quality of the good to be traded is known to the seller but not to the buyer, even if the buyer values the good more. This problem can be overcome if the seller of a high-quality good can signal its quality by taking an action that is cheaper for her to take than it would be for the seller of a low-quality good. A common action in this regard is to offer a warranty against the good's proving to be substandard (Grossman, 1981). Conversely, the buyer could screen for quality by offering a premium to any seller who agrees to provide a warranty. The signal (screen) works only if the seller is bound to honor the warranty, because a low-quality seller can offer (agree to) a worthless warranty just as cheaply as a high-quality one can. Some form of commitment is, thus, needed to implement the exchange.

1.1.3. Implementing production over time

Finally, contracts are valuable in promoting production in advance of exchange. Advance production typically increases the surplus available from exchange, but requires sinking resources in ways that may be unrecoverable if the contemplated exchange is not completed. For example, a clothing manufacturer can increase the price it receives for its products by producing them to meet the needs of its buyers, either very specifically (*e.g.*, custom-tailored suits) or only moderately so (*e.g.*, cutting them so they will be in style for a limited time only). Once the materials used to make the clothing are combined in a particular way, however, they can no longer be easily reconfigured to produce other items. In such settings, producers will be reluctant to sink such expenditures up front unless they can be assured that they will recover their costs *ex post* (see Williamson, 1975 for a seminal analysis of holdup problems).

As Katz (1996a) discusses, suppliers typically cannot capture all the surplus their pre-trade investments generate; some of this surplus will go to the buyer. Absent binding purchase commitments prior to investment, suppliers' incentives to invest will be suboptimal, possibly to the point that no investment and, so, no trade occur. Binding contracts can restore proper incentives. Conversely, buyers can also increase their surplus from exchange by making up-front investments, whether out-of-pocket (*e.g.*, buying complementary inputs) or implicit (*e.g.*, ceasing to maintain alternate sources of supply). Because such investments are often relationship specific, however, buyers will not make them unless they can be assured that the exchange price will stay sufficiently low. In some cases, the parties may be able to provide such assurance by manipulating property rights (*e.g.*, Grossman and Hart, 1986; Hart, 1989) or industrial structure (*e.g.*, Shepard, 1987) *ex ante*, but when they cannot do so cheaply, contracts could be a cost-effective alternative.

1.1.4. Limitations of contracts as commitment devices

While contracts are often useful for achieving commitment, they can be imperfect devices for doing so for some of the following reasons.

Specification costs. Because it is costly to foresee or to write down all the potential contingencies that might be relevant to the performance of the parties' contractual obligations, actual contracts are often left incomplete. Incompleteness has at least two meanings: first, the contract could simply fail to provide for certain contingencies, in which case a tribunal called upon to enforce the contract, or the parties themselves, would have to decide after the fact what to do if such contingencies arise. Second, the contract could cover all relevant contingencies, but not in as fine-tuned a manner as would be ideal insofar as the contract does not distinguish finely enough, in terms of consequent obligations, among the possible contingencies. In either event, the contract will, with positive probability, fail to assure commitment or commit the parties to a course of action that is suboptimal *ex post*.

Enforcement costs. It is never costless to hold a party to his commitment if he is inclined to try to escape it. If the contract is being enforced through the courts, for instance, lawyers must be hired and evidence assembled, and performance or damages are likely to be awarded only after some delay. Such costs make enforcement incredible when the damages from breach are relatively small; and parties can exploit this lack of credibility by holding the level of breach below the threshold necessary to provoke suit (Menell, 1983 and Priest, 1978).

Unobservable and unverifiable actions. Even if legal commitment has been established and the means for its enforcement are available, the beneficiary of a contractual promise may be unable to determine whether the promise has been kept or broken. For instance, the typical purchaser of a complex consumer product is not in a position to tell whether the product has been manufactured according to warranted specifications; at most she can observe whether the product works as expected. Even if a promisee can determine that there has been a breach, she may nevertheless be unable to demonstrate that fact to a third-party enforcer at reasonable cost. For instance, a supplier might deliver substitute goods that appear reasonably equivalent to a lay person or to a generalist court, but which the parties themselves know to be substandard. In such situations, the promisee's inability to prove that the promise has been breached renders it ineffective as a method for assuring commitment. On the other hand, as we discuss in §4.3.1, the parties can sometimes contract around the court's lack of expertise (see, *e.g.*, Hermalin and Katz, 1991 and Maskin and Tirole, 1999).

Dynamic inconsistency. In cases where the purpose of contractual commitment is to promote specific investment, the parties' incentives to stick with their deal may change after the investment has been completed (*e.g.*, Laffont and Tirole, 1988 and Aghion et al., 1994). In particular, the parties may collectively wish to modify or renegotiate their bargain. But if the parties anticipate that renegotiation will take place, it could prove impossible to induce them to undertake efficient investments *ex ante*.

The need for pre-contractual commitment. Some commitments, in order to serve their purpose, must be undertaken before the parties are in a position to engage in voluntary contracting. For example, parties may spend resources on finding contractual partners or on determining whether exchange is worthwhile. Even once an available partner and potential transaction are identified, it typically takes time and expense to negotiate terms; and commitments are often less valuable if they are delayed until bargaining is completed. (For an extreme example, consider the case of an emergency paramedic who must decide whether and how to treat an unconscious accident victim who is not carrying an insurance card.) In ongoing or repeated relationships it is possible for the parties to agree to accept liability in advance of a final bargain, but in one-shot or new relationships it is not.

The law of contracts has recognized most of these problems and has devised a variety of doctrinal arrangements to deal with them; and the succeeding sections of this chapter will discuss such arrangements in more detail. The reader should appreciate at the outset, however, that because these legal arrangements are themselves imperfect, parties will often want to use legal contracts in combination with other legal and non-legal commitment devices, such as deposits, third-party guaranties, reputational bonds, repeated dealing, mutual threats, hostage exchange, investing in altruistic preferences, and the like. The success of formal contract law, accordingly, depends importantly on how well it functions in combination with these substitute and complementary devices, and not just on how well it works in isolation.

1.2. Law & economics issues in contracting

1.2.1. Normative issues

Much normative discussion relating to contract law revolves around the issue of freedom of contract—to what extent will unregulated private contracting lead to desirable social consequences? We discuss this issue, and its relationship to standard issues in welfare economics, in §2 below. At the outset, however, it is worth observing that the dominant normative consideration here, even more so than in other fields of law and economics, is transactional efficiency. In part, this dominance follows the implicit assumption, shared by most commentators, that externalities and analogous market failures are a less significant phenomenon in this field of law than they are in, say, tort law. But the focus on efficiency also stems from the general recognition that much of contract law, putting aside specialized categories such as consumer and employment contracts, is designed for the purpose of facilitating exchange between business firms or analogous commercial entities. Such entities are motivated primarily by economic gain as opposed to nonpecuniary considerations, enter into legal obligations deliberately and at arms length, and are rational in the standard economic sense. It is thus easier to justify applying the efficiency norm to such voluntary arrangements than to the typical tort case involving persons drawn together involuntarily and outside of market institutions.

A more complete account of social welfare would consider competing normative values such as fairness, equity, etc. to the extent they affect social well-being. While there has been relatively little economic analysis of contract law in this regard, we will discuss these values insofar as they are relevant to specific analytical and doctrinal topics.

1.2.2. Positive issues

While most work on the economics of contract law has sought, at least in part, normative conclusions, there is a segment of the literature devoted to predicting and explaining how different contractual rules affect private transactions, and why contracting parties might choose one contractual device rather than others. For example, a variety of authors (*e.g.*, Joskow, 1987; Crocker and Masten, 1988; and Pirrong, 1993) have investigated the connection between the use and duration of contractual agreements and the extent of relationship-specific investments. Other authors (*e.g.*, Klein, 1980; Goldberg and Erickson, 1987; Hadfield, 1990; and Gergen, 1992) have sought to explain the common use of indefinite or open terms in otherwise clearly negotiated agreements; and still others (*e.g.*, Weinstein, 1998 and Goldberg, 1998, 2000) have sought to explain particular risk-sharing or option terms. In the field of commercial contracts, there is a vigorous literature discussing the determinants of secured lending (see, *e.g.*, Scott, 1986; Schwartz, 1989; Triantis, 1992; and Mann, 1997a, 1997b). The antitrust literature has considered whether certain contractual practices are more likely to have efficiency or anticompetitive motivations (*e.g.*, Cheung, 1969; Kenney and Klein, 1983, 2000; Crocker and

Masten, 1988; Klein and Murphy, 1988; and Masten and Snyder, 1993). While a full survey of this literature is beyond the scope of this chapter, the reader should be aware that many of the issues discussed in the later sections have been discussed empirically. At the same time, however, it also true that the empirical study of contract is relatively less developed than the theory, leaving much room for future researchers.

1.2.3. Economic versus non-economic theories of contract law

In recent years, the majority of contracts scholarship in the legal academy has reflected a methodology based on economic analysis, and most legal scholars in the field have become conversant with economic concepts such as efficiency, moral hazard, adverse selection, and the like. (Conversely, over the same period, economic theorists have increasingly come to appreciate the importance of legal concepts such as principal and agent.) Nonetheless, a considerable amount of discussion in legal circles continues to reflect alternative conceptual frameworks; and economists engaged in interdisciplinary work should be aware of these competing frameworks and their underlying assumptions. Three major competing perspectives are worth brief discussion here; we denote these as the corrective justice, liberal autonomy, and social constructivist perspectives.

Corrective justice. Corrective justice, the most longstanding of these perspectives, has intellectual roots that trace back to classical writers such as Aristotle. This perspective holds that judicial institutions are only justified in acting to redress unjust or wrongful situations. Examples of such redress in the contractual setting would include restitution of unjustly received benefits or compensation for wasted expenditures incurred in reliance on a broken promise.

The corrective justice approach seeks the restoration of some past or proper state of affairs, and thus can stand at odds with the economic approach to law, which tends to regard past gains and losses as sunk and to emphasize incentives for future behavior (*e.g.*, Easterbrook, 1984). It is less clear, however, that the corrective justice approach has any implications at all for *ex ante* analysis of legal problems; and many legal writers in this tradition (*e.g.*, Dworkin, 1980) have distinguished between the use of economics in judicial settings, which they regard as requiring decisions according to principle, and its use in legislation and contractual planning, which may legitimately be designed to promote goals of social policy or private advantage.

Liberal autonomy. The liberal or autonomy-based perspective (see, *e.g.*, Barnett, 1986) emphasizes the individual as opposed to the collective interest. From this point of view, individual rights should take priority over more general concerns of society. Such a perspective is plainly more consonant with the economic approach than is the corrective justice perspective. Standard economic measures of welfare are based on aggregates of individual utility, and under conventional assumptions of welfare economics, liberal freedoms tend to lead to desirable economic outcomes (see §2.2 below). There will be tension between the liberal and economic approaches, however, whenever market failures or transaction costs prevent the completion of efficient exchanges, as Sen (1970) has shown through social choice theory. In a contractual setting, for instance, liberal theorists might argue that obligations to which the parties have not knowingly consented may not be imposed on them even when those obligations would be both efficient and distributionally equitable. Conversely, libertarians might argue that the law should protect even anticompetitive agreements, such as price-fixing, on the basis that the conduct is voluntary and that consumers have no inherent right to trade with producers on any particular terms.

Social constructivism. The social constructivist perspective takes the position that the normative goals of society—and in the view of many writers, its descriptive categories as well—are determined by collective and ongoing deliberation among its citizens, and accordingly, that the main goal of legal institutions should be to provide adequate opportunity for such decision making. So defined, this perspective can be seen to encompass a variety of more specific normative positions, including those of writers who emphasize civic virtue (Kronman, 1987), democratic self-government (Pildes and Anderson, 1990), an egalitarian distribution of political and economic power (Kennedy, 1982), or the primacy of particular substantive values, which economists would call merit goods (Radin, 1987). On such views, the constituted citizenry may well decide to pursue economic goals such as efficiency, but may, with equal legitimacy, decide to pursue other procedural or substantive goals. For example, if the citizenry decided the well-being of local manufacturers were sufficiently important to outweigh productive efficiency or economic liberty, this decision would legitimize restricting interregional trade. It follows that there is no particular reason why individualist institutions, like contract or the market, are more legitimate venues for such decision making than collective or political ones.

Relationship between economic and non-economic theories. These three rival perspectives can and often do combine to provide overlapping arguments against the use of economic analysis in contract law. For example, the most prominent alternative theory of contract law to be put forward in recent years, the so-called "will theory" (*e.g.*, Fried, 1981), holds that promises ought to be kept for their own sake, in part because promise-breaking is a deontological wrong that needs to be rectified as a matter of corrective justice, and in part because enforcing promises is necessary to respect the autonomy of the promisee (and arguably the promisor as well). In response to such arguments, more economically oriented writers (*e.g.*, Craswell, 1989a and Shavell, 1991) have replied that most economic analysis of contract law is aimed at filling the gaps in incomplete agreements and setting default rules that operate when the parties have expressed no preference regarding a particular issue. If we accept that most business contracts have economic purposes, economic analysis will help those interpreting contracts to better implement the parties' will.

The will theorists might offer, as a rejoinder, the observation that there is a difference between what most contracting parties subjectively understand when they make and receive promises, and what would be most efficient for them to do. For instance, most people, even those with substantial business experience, intuitively understand promises to bear inherent moral force and believe that the mere fact that it turns out to be suboptimal to carry them out should not count as an excuse for non-performance. If this claim is true—and it should be plain that it is an empirical claim—then will theorists (*e.g.*, Charny, 1991) would argue that it violates the parties' autonomy and dignity for the state to enforce the efficient bargain, which they perhaps should have made, rather than the actual arrangements that they did make.

An obvious response for economists to such non-economic considerations is to incorporate them into a larger welfare analysis in which the relevant non-economic values are interpreted as arguments to be traded off against one another in a Bergson-type social welfare function (SWF). Shavell and Kaplow (2002) have presented such an analysis at length, but have achieved mixed success in persuading non-economically-oriented legal scholars of its merits. Observe, in this regard, that the three main categories of non-economic theories vary in their compatibility with the SWF approach. Pure corrective-justice theorists would reject the SWF on the grounds that legally appropriate actions can only be justified with reference to right and wrong, and never with reference to consequences. Liberal autonomy theorists tend to be somewhat more accepting of the SWF approach, but they would reject the idea that rights could be traded off against each other or against economic values (formally, this is equivalent to adopting a lexicographical preference ordering for the SWF). Social constructivists would view the whole SWF approach as logically circular, since they view the social values that would underlie any SWF as endogenously determined by political and cultural processes of which economic policy discussions are an integral part. In their view, starting with a SWF and then attempting to maximize it is putting the cart before the horse.

Most importantly, economists who study legal institutions should recognize that lawyers and legal commentators often employ different methodologies than they do. In particular, legal scholars do not typically draw the same sharp distinction between positive and normative analysis as economists do. On the contrary, many seemingly descriptive statements made by lawyers or appearing in legal texts are understood in context to carry significant normative overtones, and *vice versa*. Non-lawyers who are insufficiently appreciative of the mixed nature of such discourse may miss an important part of what is being said, and may be led to model the phenomena under discussion in an incorrect or misleading way (Katz, 1996a). Similarly, lawyers' customary method of reasoning inductively from individual cases, rather than deductively from general principles, makes many lawyers reluctant to accept some of the standard methodological practices of economists, including formal modeling and the use of statistical aggregation. When presenting positive arguments or analyses to legal audiences, accordingly, it is generally necessary to lay out one's methodological assumptions explicitly and at the outset. Otherwise, one risks being misunderstood by those lawyers who are accustomed to assuming a normative subtext or an individualist perspective whenever legal topics are being discussed.

1.3. What this chapter is not

1.3.1. A guide to contract theory

Not surprisingly, there is a strong link between the law and economics of contracts and the economic sub-discipline of contract theory.[1] Contract theory provides a framework to analyze the scope and limits of what contracts can accomplish, at least at a theoretical level. It is, however, beyond the scope of this chapter to offer a comprehensive guide to contract theory. To be sure, some contract theory will be discussed as appropriate within the context of the issues discussed below; but no one should mistake such a scatter-shot approach for an attempt at a systematic treatment of the subject.

For readers interested in contract theory, there are a number of excellent introductions. Most new graduate microeconomics texts devote at least a chapter to the subject (see, *e.g.*, Kreps, 1990; Mas-Colell et al., 1995; or Varian, 1992). Good book-length introductions are Laffont and Martimont (2002) and Bolton and Dewatripont (2005). For those looking for cheap—in fact free—introductions, one of us (Hermalin) hosts two introductory manuscripts (Caillaud and Hermalin, 2000a, 2000b) on his web page.

1.3.2. A guide to the law of contracts

While this chapter discusses most of the important economic issues relating to the law of contracts, it does not attempt to present a survey of major legal doctrines in the field. It is worth noting, as a general observation, that many legal doctrines may seem indistinguishable from an economic viewpoint. Nevertheless, they vary in their specific content and their differences are of practical importance to lawyers and clients. For example, the doctrines of mistake, impossibility, and frustration of purpose all excuse contractual liability in extreme or unexpected circumstances and thus allocate risk to the recipient of a contractual promise, but their legal application varies.

Economists interacting with lawyers or pursuing research in contract law, accordingly, should be aware of distinctions such as these and of the associated views of major legal authorities. The most useful general treatise on US contract law is probably Farnsworth (2004), available in both one- and three-volume editions; a shorter introduction to the subject can be found in Chirelstein (2006). For statutory material applicable to US contracts for the sale of goods, the best source is White and Summers (2000).

Less attention has been paid by economists to non-US legal systems, making them a potentially fruitful source for future research; here leading English-language treatises include Honnold (1999) and Schlechtreim (1998) on international sales contracts,

[1] Whether this link is a valuable one is, however, a matter of debate. See, for instance, the *Yale Law Journal* debate among E. Posner (2003a), Ayres (2003), and Craswell (2003).

Bonell (1997) on other international commercial contracts, Atiyah (2003) and Treitel (2003) on English contract law, and Lando and Beale (2002) on the contract law of the civil law systems of continental Europe. In this chapter, however, most discussions of legal doctrine will be restricted to US law.

1.4. Organization of the chapter

The remainder of this chapter is divided into four sections corresponding to the major conceptual divisions of contract doctrine. Section 2, entitled "Freedom of contract," discusses the scope of private parties' power to create binding contractual obligations. It analyzes the major doctrines that govern which bargains will be recognized and enforced by state legal institutions, and which parties are empowered to create enforceable bargains. It also considers how doctrinal limits on freedom of contract correspond to the economic criteria for determining whether decentralized trade will lead to optimal welfare outcomes.

Section 3, entitled "Formation of contracts," discusses the extensive body of legal doctrine that governs the procedural mechanics of exchange, as well as the rules that govern the parties' obligations before they enter into exchange. These formal rules, by attaching consequences to the various acts and omissions that bargainers can choose from in searching for and negotiating with potential contractual partners, affect parties' incentives to make and to respond to offers, to delay, to bluff, and to communicate with one another in the first place.

Section 4, entitled "Interpretation of contracts: Contractual incompleteness," discusses the problems that arise when it is unclear whether those parties who are empowered to create binding contracts have actually done so, and if they have, what specific obligations they have created. Recent work in microeconomic theory has also been concerned with this problem, especially as it relates to the ability of third-party enforcers to verify the parties' bargain. From a legal perspective, the problem is governed by the various doctrines dealing with contract interpretation, and this section shows how the legal rules in this area affect and respond to the economic problem of incomplete contracts.

Section 5, entitled "Enforcement of contracts," discusses how the foregoing rules and institutions are translated into effective costs and benefits that can motivate parties to comply with their obligations and to insure against others' lack of compliance.

Finally, the concluding section offers some overall perspectives on the entire discussion, relates its main points to analogous or complementary doctrines in related fields of law, and offers some speculations regarding the path of future legal and economic developments in the area.

2. Freedom of contract

The threshold issue in any discussion of contract law is freedom of contract—the extent to which the law sanctions the use of contracts as a commitment device. No legal system enforces all voluntary private agreements, but in the US and other industrial democracies, most contracts that support legitimate economic exchange are at least presumptively enforceable. Still, the limits of freedom of contract vary among Western countries and are an important element of regulatory policy. This section, accordingly, analyzes and evaluates those limits in economic terms. Subsection 2.1 defines the scope of the issues included under the framework of contractual freedom; subsection 2.2 reviews the presumptive economic case in favor of freedom of contract; subsections 2.3 and 2.4 discuss the main arguments, economic and otherwise, that are typically used to justify limits on private contracting; and subsection 2.5 outlines the major doctrinal limitations on freedom of contract that are in force in the US and in related systems, and relates those doctrines to the economic arguments set out in the prior subsections.

2.1. Freedom of contract defined

2.1.1. State regulation versus state enforcement

The concept of contractual freedom encompasses a number of distinct considerations. One important distinction is between negative and affirmative government sanction: are the parties permitted to enter into a given contract versus will the law enforce it? These two questions are not equivalent: there are many agreements that cannot be enforced in the courts but that can still be useful as commitment devices if the parties can manage to implement them privately. For instance, prior to the passage of the Sherman Act, contracts in restraint of trade were unenforceable under US common law; this reduced their incidence, but not to the point of elimination. Under modern antitrust statutes, in contrast, government disapproval of anticompetitive conduct goes beyond non-enforcement to include active interference through civil liability and, in some cases, criminal prosecution.

In the remainder of this section, accordingly, we focus on those limits on contractual power that are motivated by regulatory concerns that the agreement itself is socially undesirable for reasons of inefficiency, inequity, and other substantive objections.

2.1.2. Positive versus negative contractual freedom

The freedom to enter into contractual liability would be rather less meaningful were it not accompanied by the complementary freedom to avoid liability for contracts into which one does not wish to enter. In general, this negative freedom applies to most types of contractual obligations, but not all. In traditional common law, for instance, some businesses (*e.g.*, mills, ferryboats, railroads, and the like) are designated as common carriers and are obliged to enter into exchange on standard terms with anyone who wishes. Modern statutes have expanded such duties in a variety of ways: for instance, the essential-facilities doctrine in antitrust law requires vertically integrated firms to make certain stages of production available on a contractual basis to their non-integrated competitors; and anti-discrimination laws require businesses to deal with customers and suppliers on an equal basis without regard to race, religion, or the like.

Additionally, some rules of tort and property law have the effect of requiring rightholders, under some circumstances, to transfer their entitlements to persons in need or to the general public. For instance, the doctrine of eminent domain requires landowners to convey their property to the state when required for an appropriately defined public use. (E.g., in the US, the government could be constitutionally obliged to pay just compensation when it takes private property.) Similarly, the tort doctrine of necessity allows parties in dire need to make use of others' property when voluntary contracting is not feasible (as with an endangered hiker who breaks into an empty cabin to find food or shelter) so long as compensation is paid *ex post*.[2] Such doctrines are typically conceptualized by legal theorists as internal limits on the underlying entitlement at issue rather than as restrictions on contractual freedom; however, they usually amount in practice to restrictions on contractual freedom because of the infeasibility of contracting around them. One could theoretically imagine contracting in advance with the government not to exercise its right of eminent domain, but such contracts are rare and it is doubtful whether they are immune to abrogation by subsequent governments.

Finally, some doctrines of contract law impose promissory liability even when the promisor has not actually intended to enter into a contractual exchange. The doctrine of promissory estoppel, for instance, holds parties to promises on which a reasonable person would foreseeably rely, at least to the extent necessary to protect the promisee's reliance; and the doctrine of trade usage holds parties to contracts that experienced market participants would view as legally enforceable under similar circumstances, even if the parties had not themselves subjectively intended to be bound in the particular instance. For the most part, such doctrines provide rules of interpretation rather than of substantive contractual freedom, in that it is possible to avoid liability by being sufficiently explicit in one's communications. In some cases, however, the law does not allow one to disclaim liability for one's representations or promises; for instance, commercially sophisticated parties dealing with less sophisticated counterparts can find themselves bound to statements made to the less experienced party even if sophisticated persons would understand such statements do not entail legal obligation. Such restrictions on disclaimers are usually motivated by considerations of market failure, transaction cost, or distribution, as discussed below.

2.1.3. Mandatory versus default terms

Finally, as the example of promissory estoppel illustrates, it is useful to distinguish between mandatory contract terms, which the parties are not legally free to change, and

[2] Modern legal theorists, following Calabresi and Melamed (1972), use the concept of property and liability rules, further discussed in this *Handbook*, to identify situations in which such involuntary exchanges are authorized by law: entitlements subject to such imposition are said to be protected by a liability rule, while entitlements that are immune from such imposition are said to be protected by a property rule.

default terms, which the parties are theoretically free to change but which govern the contract to the extent the parties are silent. An example of the former would be the constitutional prohibition on involuntary servitude, which, for instance, prevents people from pledging their future labor for long periods, even when they might find wish to do so, such as to provide collateral for a loan. An example of the latter would be the various warranties of quality that are implied under current US law in contracts for the sale of goods. Sellers are generally permitted to disclaim such warranties, subject to some limits imposed by consumer-protection and product-liability law, by following the requisite procedures, which usually require some specific notification to the buyer.

Because it is costly to write complete contracts, all systems of contract law must provide default terms to cover the issues over which the parties do not specifically bargain (see §4.4.1 below). The regulatory effects of default terms, however, are bounded by the costs of contracting around them. It could be a reasonable approximation to ignore such effects in many instances, especially in the commercial setting where parties are sophisticated and have access to legal advice. In some cases, however, re-contracting costs are substantial and the choice of default rule will have the effect of privileging one outcome over others. For example, in mass transactions in which parties communicate through standard forms, it is impractical to reconcile all the discrepancies between the various forms; and under modern doctrine, the legal default applies to all issues on which the forms do not agree. This often has the consequence of providing broad product warranties and leaving the seller open to liability for damages following breach, even if the seller attempted to disclaim such liability.

As with the decision to withhold enforcement, the provision of default rules may be motivated either by regulatory purposes or by the desire to conserve on transaction costs. In situations where there is some doubt about whether a specific contract term should be discouraged, for example, a default rule supplies some deterrent effect while still allowing parties whose gain from the term is sufficiently high to opt out of the default at a price. For example, Camerer et al. (2003) advocate using default rules as a relatively libertarian method of regulating against poor decisions caused by bounded rationality. Or, as Ayres and Gertner (1989, 1992) have suggested, and as we discuss further in §4.4.1 below, a default rule may be employed as a screen to induce parties to reveal private information that might be relevant to *ex post* interpretation or to the *ex ante* decision whether to enter into exchange.

2.2. The economic case for freedom of contract

2.2.1. Welfare economics

An economic case for or against freedom of contract is based on the consequent welfare implications. In this section, we briefly review what a welfare analysis of markets suggests about freedom of contract.

Economists typically use two welfare criteria. One, known as Pareto efficiency, evaluates a proposed allocation among a set of actors by asking whether there exists a second allocation that (i) none of the actors prefer less than the proposed allocation and (ii) at least one of the actors actually prefers to the proposed allocation. If such a second allocation exists, the proposed allocation is deemed inefficient (alternatively, Pareto inferior or Pareto dominated). The second allocation in this case is deemed Pareto superior. If no such second allocation exists, the proposed allocation is deemed efficient.

While the Pareto criterion is useful for ruling *out* undesirable allocations, it doesn't always serve as a useful guide for selecting a desirable allocation. For instance, it offers no guidance as to who should receive an indivisible object; any allocation other than throwing it away is Pareto efficient because a switch to another allocation would not be favored by the party losing the object.

An alternative welfare measure is to consider a function that aggregates, in some way, the preferences of the actors in question. A full discussion of welfare functions is beyond the scope of this chapter.[3] We will limit attention to the utilitarian welfare function, $W = \sum_{i \in \mathcal{I}} u_i$, where $\mathcal{I}$ is an index set over the actors and u_i is the utility the ith actor enjoys from the proposed allocation (if there is a stochastic aspect to the allocation, u_i should be understood to be i's *expected* utility). Any allocation that maximizes social welfare, W, must be Pareto efficient;[4] but Pareto efficiency does not necessarily imply that social welfare is maximized. For instance, as noted, any allocation of an indivisible object (other than throwing it out) is Pareto efficient, but only the allocation that awards it to the person who values it most is welfare maximizing.

A stronger connection between the two welfare criteria can be achieved if one accepts the existence of a transferable good (typically taken to be money). Now the losers from moving to a social welfare-maximizing allocation can receive payments from the winners as compensation. If preferences can be captured by a quasi-linear utility function (*i.e.*, of the form $u + y$, where y is money and u is utility from other goods), then an allocation is Pareto efficient if and only if it maximizes welfare.[5] For this reason, economists are generally satisfied with social welfare (total surplus) as an appropriate

[3] The interested reader is directed to Chapter 22 of Mas-Colell et al. (1995) or Chapter 6 of Laffont (1988). Also see Arrow's (1963) classic book.

[4] Proof: Suppose not. Then, although the allocation **a** maximizes W, there is another allocation $\tilde{\mathbf{a}}$ that Pareto dominates **a**. Let u_i and $\tilde{u}_i$ denote the utilities under the allocations **a** and $\tilde{\mathbf{a}}$ respectively. By the definition of Pareto dominance, $\tilde{u}_i \geq u_i$ for all i and there is at least one i such that $\tilde{u}_i > u_i$. But, then,

$$\sum_{i \in \mathcal{I}} \tilde{u}_i > \sum_{i \in \mathcal{I}} u_i,$$

which contradicts the assertion that **a** maximizes social welfare.

[5] Proof: Footnote 4 *supra* established that welfare maximization implies Pareto optimality. Consider a Pareto optimal allocation of real goods and money. Let u_i and y_i denote the utility components under this Pareto optimal allocation. Suppose this allocation does not maximize welfare. Then, there exists another allocation of real goods, with utility components u_i^* such that

$$\sum_{i \in \mathcal{I}} u_i^* > \sum_{i \in \mathcal{I}} u_i.$$

welfare standard when transfers are feasible.[6] Note that this analysis relies on the marginal utility of the transferable good being constant across individuals, so that we are still seeking to maximize $\sum_{i\in\mathcal{I}} u_i$; transfers among the actors are irrelevant to maximizing welfare because the benefit one actor gets from receiving a dollar is completely offset by the cost another incurs transferring that dollar.

Competitive markets are typically seen as doing well with respect to the Pareto efficiency criterion. Under somewhat stringent conditions—in particular, (i) that complete markets for all commodities (including commodities such as clean air and water) exist, (ii) that no actor has market power (*i.e.*, acts as a price setter rather than a price taker), and (iii) symmetric information—a general equilibrium of the economy will be Pareto efficient (see, *e.g.*, §6.3 of Debreu, 1959). This result is known as the First Welfare Theorem.[7]

The First Welfare Theorem applies to the economy as a whole. If the entire economy is an Arrow–Debreu economy and in equilibrium, then any particular market within the economy must be "efficient," insofar as any change in it changes at least part of the overall allocation and no other overall allocation is Pareto superior. If, however, one doubts the Arrow–Debreu model adequately models the *entire* economy, one can still ask, in certain circumstances, whether a particular market in it is achieving an efficient allocation.

A single competitive market (*i.e.*, a market in which neither buyers nor sellers exercise market power) will achieve an equilibrium at the price that equates demand and

Because the "u_i" allocation is Pareto efficient, it cannot be that $u_i^* \geq u_i$ for all i; that is, a change in allocation must create "losers," for whom $u_i < u_i^*$. Let $\mathcal{L}$ be the set of losers and L be the number of elements in $\mathcal{L}$. As just noted, $L > 0$. Similarly, there must be winners (*i.e.*, those for whom $u_i^* \geq u_i$). Let $\mathcal{W}$ be the set of winners. For losers, define $\tau_i = u_i - u_i^*$. For winners, define $t_i = u_i^* - u_i$. Because the "u_i^*" allocation is welfare maximizing,

$$\sum_{i\in\mathcal{W}} t_i - \sum_{i\in\mathcal{L}} \tau_i \equiv G > 0.$$

Finally, consider the allocation of real goods that produces utilities u_i^*, allocates $y_i - t_i$ in money to each winner and $y_i + \tau_i + G/L$ to each loser (note the additional transfers sum to zero, hence are feasible). For a winner, $u_i^* + y_i - t_i = u_i + y_i$, so winners are indifferent. For a loser, $u_i^* + y_i + \tau_i + G/L = u_i + y_i + G/L$, so losers are strictly better off. But, then, we have an allocation (including money) that Pareto dominates our original allocation, contradicting its Pareto optimality. The result follows by contradiction.

[6] The Kaldor-Hicks criterion is even more flexible: An allocation, $\boldsymbol{\alpha}$, over real goods is a Kaldor-Hicks efficient allocation if there is no other allocation over real goods $\boldsymbol{\alpha}'$ and no profile of transfers, $\mathbf{t}$, such that the overall allocation $(\boldsymbol{\alpha}', \mathbf{t})$ Pareto dominates the overall allocation $(\boldsymbol{\alpha}, \mathbf{0})$, where $\mathbf{0}$ means no transfers are made. It is sufficient that $\mathbf{t}$ exist—whether or not these transfers are made—for $\boldsymbol{\alpha}'$ to dominate $\boldsymbol{\alpha}$ according to the Kaldor-Hicks criterion. If all individuals have utility $u_i + y_i$, where u_i is an individual's utility over real goods and y_i is his allocation of money (transfer), then an allocation of real goods is Kaldor-Hicks efficient if and only if it maximizes $\sum_{i\in\mathcal{I}} u_i$. See Chapter IV of Arrow (1963) for details.

[7] There is also a second Welfare Theorem that has to do with the ability of prices to serve as appropriate incentive devices in a competitive (Arrow–Debreu) economy; that is, any Pareto efficient allocation can be supported as a general equilibrium of the economy by selecting the appropriate prices (see §6.4 of Debreu for details).

supply. Under normal assumptions, this equilibrium is unique. Assume the demand curve is a good approximation for the social marginal benefit curve (*i.e.*, there are essentially no positive externalities from the good in question and income effects are *de minimis*).[8] Assume, further, that the supply curve is a good approximation for the social marginal cost curve (*i.e.*, there are essentially no negative externalities from the good in question). Then, as is well known, welfare is the area beneath demand and above supply from 0 to the number of units traded. This area—that is, total welfare—is maximized by exchanging the quantity that corresponds to the intersection of supply and demand. But this quantity is precisely the quantity that will be exchanged in competitive equilibrium—the competitive equilibrium maximizes total welfare and, thus, achieves a Pareto-efficient allocation.

2.2.2. *Theoretical justifications for freedom of contract*

The welfare-theoretic arguments of the previous subsection rely on the assumption of competitive markets. Many writers, however, have had the intuition that freedom of contract is desirable much more generally.

In the law and economics literature, this intuition is most prominently associated with the work of Ronald Coase and his widely-cited "Coase Theorem" (Coase, 1960). Despite its formal sounding name, the Coase theorem is not a theorem in the traditional sense (nor did Coase suggest it was). Indeed, as Medema and Zerbe (2000) point out, there is not even an agreed upon statement of it. We offer the following version:

THEOREM 1 (Coase Theorem). *Consider a bilateral contracting situation in which (i) the parties are rational with respect to their individual self-interests; in which (ii) the parties can agree on any contract without incurring transaction costs; and in which (iii) the parties' utilities are additively separable over the allocation of real goods and monetary transfers (i.e., are of the form $u_i + y_i$, where u_i is party i's utility from the allocation of the real goods and y_i is his or her net transfer). Then the allocation of real goods after contracting will maximize total welfare regardless of the initial allocation of real goods.*

In this formulation, the Coase theorem is a true theorem with the following proof: Let $\boldsymbol{\alpha}$ be any real-good allocation that does *not* maximize welfare. We need to show that no such $\boldsymbol{\alpha}$ will be implemented via contracting. Let $\boldsymbol{\alpha}^*$ be a welfare-maximizing allocation. Clearly, if $u_i(\boldsymbol{\alpha}^*) > u_i(\boldsymbol{\alpha})$ for both parties i, then $\boldsymbol{\alpha}$ will not be implemented—because both parties act rationally and they can costlessly contract, they won't settle for $\boldsymbol{\alpha}$. Consider the only other possibility: $u_i(\boldsymbol{\alpha}^*) > u_i(\boldsymbol{\alpha})$ for one party and $u_j(\boldsymbol{\alpha}^*) \leq u_j(\boldsymbol{\alpha})$ for the other. Let $\tau_M = u_i(\boldsymbol{\alpha}^*) - u_i(\boldsymbol{\alpha})$ and $\tau_m = u_j(\boldsymbol{\alpha}) - u_j(\boldsymbol{\alpha}^*)$. The optimality of $\boldsymbol{\alpha}^*$

[8] For a detailed discussion of measuring consumer benefit and the consequence of income effects for such measurements see Chapter 10 of Varian (1992). Also see Willig (1976).

entails $\tau_M > \tau_m$. Pick any transfer τ such that $\tau_m < \tau < \tau_M$. Then a contract that selected allocation $\boldsymbol{\alpha}^*$ and had i transfer an additional τ to j would be preferred by both parties to a contract that implemented allocation $\boldsymbol{\alpha}$. Costless contracting and the parties' rationality thus rule out $\boldsymbol{\alpha}$ being the contracted-for allocation.

COROLLARY 1. *Under the assumptions of the Coase Theorem, interference or restrictions on the contract the parties sign cannot increase total welfare; that is, under these assumptions, there should be freedom of contract if the only welfare issue is the total welfare of the parties to the contract.*

How strong a case the Coase Theorem makes for freedom of contract depends on the appeal of its assumptions. Consider, first, assumption (iii); that the parties have quasi-linear utility functions. That assumption is common to most welfare analyses of partial-equilibrium settings. Although the assumption can frequently be justified, it is not always.[9] Suppose, for instance, that one party's utility is $u + y$ if $y \geq 0$, but $-\infty$ if $y < 0$. An interpretation is that this party simply cannot survive if made to make transfers. Now the Coase result could fail to hold if this party is the one who must transfer to ensure a welfare-maximizing outcome. On the other hand, if this party is the recipient of transfers, then welfare-maximization would continue to hold.

As Medema and Zerbe (2000) note, the Coase Theorem has two conclusions. One is an efficiency conclusion—private contracting will lead to a welfare-maximizing solution. The other is an invariance conclusion—the initial allocation is immaterial for whether a welfare-maximizing solution is reached. As just discussed, relaxing condition (iii) can undermine both conclusions, but, in a sense, it primarily undermines the invariance result. Efficiency, as judged by the Pareto criterion, will generally still be attained:

THEOREM 2 (Modified Coase Theorem I). *Consider a bilateral contracting situation in which (i) the parties are rational with respect to their individual self-interests, but are not mean-spirited; and in which (ii) the parties can agree on any contract without incurring transaction costs. Then the allocation after contracting will be Pareto efficient regardless of the initial allocation.*

PROOF. We need to show that the parties will never settle on an allocation, **a**, that is Pareto dominated (note, now, an allocation may include the transferable good). Consider such an allocation. By definition, there exists at least one other allocation, $\mathbf{a}^*$, such that

$$u_i(\mathbf{a}^*) \geq u_i(\mathbf{a}) \tag{1}$$

for both i and such that $u_i(\mathbf{a}^*) > u_i(\mathbf{a})$ for at least one i. Case 1: The inequality in expression (1) is strict for both i. Then **a** will not be implemented—both parties act

[9] Willig (1976) provides justifications for assuming quasi-linear utility that are applicable in many contexts.

rationally and they can costlessly contract, so they won't settle for **a**. Case 2: (1) is an equality for one i. Then that i will refuse to implement $\mathbf{a}^*$ over **a** only if he is mean-spirited, which we have assumed he isn't. Hence, a Pareto-dominated allocation will not be implemented. □

The modicum of altruism in the modified Coase Theorem—that the parties not be mean spirited—is unnecessary if we assume a strict-tradeoff condition:

CONDITION 1 (Strict Tradeoff). *Let $\mathcal{A}$ be the set of feasible allocations and $\mathbf{a}_1$ and $\mathbf{a}_0$ be two elements of $\mathcal{A}$. Then, if $u_1(\mathbf{a}_1) \geq u_1(\mathbf{a}_0)$ and $u_2(\mathbf{a}_1) \geq u_2(\mathbf{a}_0)$ with at least one inequality holding strictly, there exists an $\hat{\mathbf{a}} \in \mathcal{A}$ such that $u_1(\hat{\mathbf{a}}) > u_1(\mathbf{a}_0)$ and $u_2(\hat{\mathbf{a}}) > u_2(\mathbf{a}_0)$.*

The strict-tradeoff condition entails that if **a** is a Pareto-dominated allocation, then there exists a feasible allocation that strictly Pareto dominates **a**; in which case, we need only Case 1 of the proof of the modified Coase Theorem and we can dispense with the case that relied on no mean-spirited behavior:

THEOREM 3 (Modified Coase Theorem II). *Consider a bilateral contracting situation in which (i) the parties are rational with respect to their individual self-interests; in which (ii) the parties can agree on any contract without incurring transaction costs; and in which (iii) the set of feasible allocations satisfies the strict-tradeoff condition. Then the allocation after contracting will be Pareto efficient regardless of the initial allocation.*

With respect to freedom of contract, we have

COROLLARY 2. *Under the assumptions of the modified Coase Theorems, interference or restrictions on the contract the parties sign cannot increase the Pareto efficiency of the contracted-for outcome; that is, there should be freedom of contract if the only welfare issue is the efficiency of the outcome achieved by the contract from the perspective of the parties to the contract.*

Clearly, all the Coase theorems rely on the rationality assumption. We explore the consequences of relaxing this assumption later (see §2.3.4 and §4.2.1).

The no-transactions-cost assumption is also important. In the proofs, it serves to guarantee that the parties will not agree to a non-optimal contract because they can, without cost or impediment, choose an optimal contract instead. Relaxing this assumption would, thus, seem to have the potential for undermining the conclusions of these theorems—a point made by a number of authors (see, *e.g.*, Farrell, 1987b).

As Farrell notes, the economic literature on bargaining is generally sanguine about the prospects of an efficient outcome when the parties bargain under symmetric information. The literature is much more pessimistic, however, when they bargain under

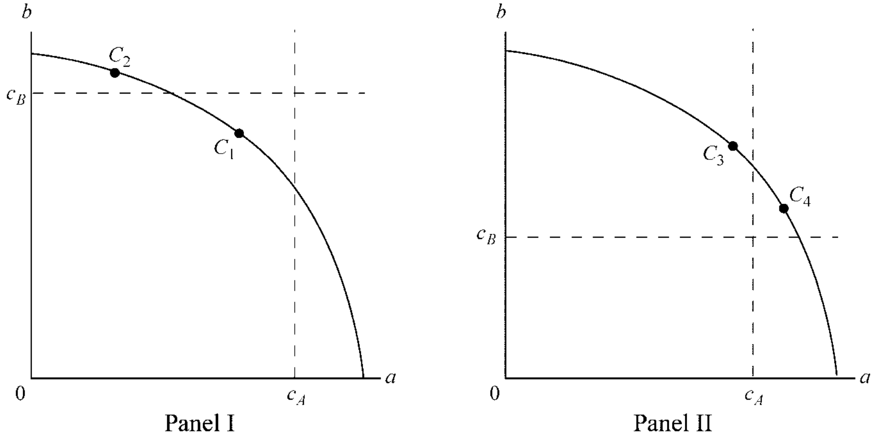

Figure 2. Consequence of transactions costs on bargaining outcomes under symmetric information.

asymmetric information.[10] We discuss the consequences of bargaining under asymmetric information in §2.3.2. The rest of this section considers what can still potentially go wrong under symmetric information and concludes with Coase-theorem-like propositions that account for the potential of costly bargaining.

The potential consequences of transactions costs with symmetric information are illustrated in Figure 2. Suppose two parties, A and B, wish to enter into a contract. To do so, however, they must expend costs c_A and c_B, respectively. Once those costs are sunk (*e.g.*, lawyers are retained), the parties bargain over the contract terms. Let the payoffs, in money, be a and b, respectively. Because bargaining is conducted under symmetric information, theory predicts that the outcome will generally be efficient (see *e.g.*, Farrell, 1987b, or Sutton, 1986);[11] that is, the contract will be on the Pareto frontier, which is shown in the two panels of Figure 2 as a solid curve.

The first problem, illustrated in Panel I, is that c_A and c_B are so large that one or both sides prefer not to enter into negotiations. For instance, if bargaining would result in contract C_1, then neither side would be willing to negotiate; at C_1, we have $a < c_A$ and $b < c_B$. Moreover, while there are contracts that, if adopted, would induce one side to participate (*e.g.*, C_2, for which $b > c_B$), there is no contract that would induce both to participate.

The problem illustrated in Panel I can serve to justify *default* contracts. That is, if the law stipulated that, absent a contract, the implicit contract was C_1, then that would clearly be an improvement. Observe that this improvement stems from two factors. First, the parties get to the Pareto frontier when they otherwise wouldn't. Second, they avoid

[10] See the survey on bargaining by Sutton (1986).

[11] Although efficiency is expected, it should be noted that one can construct perverse bargaining games that don't achieve efficiency even under symmetric information; see Hermalin and Katz (1993).

the expenditures c_A and c_B. Note, however, that this analysis does *not* rely on C_1 being *mandatory*; this is not an argument for mandatory contracts.

A related problem, and one that *could* justify mandatory contracts, is illustrated in Panel II. Suppose that bargaining would lead the parties to contract C_3. Because $a < c_A$ at C_3, party A would refuse to enter into negotiations. This is undesirable, especially as there exist contracts, such as C_4, that, were they the outcome of bargaining, would induce both parties to negotiate. Now there is scope for limitations on contracts. If terms were limited, so that the only contracts that could be considered were on the arc segment between the dotted lines (*i.e.*, the segment containing C_4), then both A and B would be willing to enter into negotiations and they would get an outcome on the Pareto frontier.

Of course, one may wonder how the law would know that C_1 is a good default contract or to limit contracts to those in the neighborhood of C_4. If the law incurs costs arriving at default or mandatory contract terms, or lacks the information to design optimal contract, then it could be better to leave things in private hands. In other words, while it is true that restrictions on private contracts can possibly enhance efficiency when the private parties incur transactions costs, one must assess that observation in light of real-life limitations on what the legal system can do and the cost at which it can do it.

Before leaving the issue of transactions costs, it is worth considering how far we can go in relaxing the no-transactions-cost assumption and still establish a Coase-like case for freedom of contract. Our objective is to have a precise statement for a result of the form *if bargaining leads the parties to a second-best efficient*[12] *contract and does so in a way that minimizes bargaining costs, then the legal system can do no better than to leave the private parties' choice of contract unrestricted.*

To establish such a result, it is helpful to switch from working with allocations to working directly with contracts. Let C denote an arbitrary contract. Contract terms are assumed not only to fix the allocation of real goods (possibly contingently), but also the allocation of transfers (possibly contingently) between the parties. Let $U_i(C)$ denote party i's utility (possibly expected) should the parties agree to contract C.

We can restate the strict tradeoff condition as

CONDITION 2 (Strict Tradeoff). *Let $\mathcal{C}$ be the set of feasible contracts and consider any two contracts C_0 and C_1 in $\mathcal{C}$. If $U_1(C_1) \geq U_1(C_0)$ and $U_2(C_1) \geq U_2(C_0)$, with at least one inequality holding strictly, then there exists a contract $C_2 \in \mathcal{C}$ such that $U_1(C_2) > U_1(C_0)$ and $U_2(C_2) > U_2(C_0)$.*

Given the Strict Tradeoff condition, the following Coase-like theorem can be established:

[12] Second-best efficiency refers to the optimal outcome taking into account the informational constraints faced by the parties. For instance, in the standard hidden-action agency problem, the first-best outcome would entail the agent expending some ideal level of effort for a flat wage. But that outcome is typically infeasible once the constraint that the agent's action is hidden from the principal is taken into account; the second-best solution in such cases typically requires the agent to bear risk, which is inefficient from a first-best perspective.

PROPOSITION 1 (Hermalin and Katz, 1993). *Suppose the two parties to a contract are symmetrically informed prior to and during bargaining and that bargaining consists of alternating offers. Assume the costs of delay in achieving an agreement are due to discounting. Assume that the set of possible contracts, $\mathcal{C}$, is invariant over rounds of bargaining and that the set $\{(U_1(C), U_2(C)) | C \in \mathcal{C}\}$ is convex and compact. Assume that there is at least one $C \in \mathcal{C}$ that each party strictly prefers to no agreement (the status quo). Finally assume either the Strict Tradeoff condition is satisfied or both parties are risk neutral and non-contingent (lump-sum) transfers in any amount are feasible. Then there is an essentially unique subgame-perfect equilibrium in which bargaining ends in the first round with agreement on a Pareto-efficient contract.*

The qualifier "essentially unique" captures the fact that there could be more than one contract that yields the unique equilibrium utility levels.

PROOF. Follows from Lemma A2 and Proposition 4 of Hermalin and Katz (1993). □

Because bargaining ends in the first round, there are no transactions costs on the equilibrium path. Hence, there is no possibility of increasing efficiency by reducing transactions costs. Moreover, because the resulting contract is efficient, there is no scope for increasing efficiency with respect to the contract chosen. In sum, under the assumptions of Proposition 1, there is no gain to be had from restricting private contracts; that is, Proposition 1 makes a case for freedom of contract.

On the other hand, Proposition 1 relies on a number of assumptions. Fortunately, some of these can be relaxed if one is willing to impose refinements on the subgame-perfect solution concept as applied to bargaining games. Specifically, we wish to rule out the possibility that the parties fear proposing efficient contracts because the equilibrium specifies a continuation game following the proposal of an efficient contract by party i that is unfavorable to party i (see Hermalin and Katz, 1993 for an example of such a perverse game). To that end, consider the following equilibrium refinements:

CONDITION 3 (Monotone Acceptance Condition). *Suppose that, at some date (round) t, a party would accept an offer of contract C_0. Suppose that another contract, C_1, would yield that party a higher expected utility level. Then, he or she will also accept contract C_1 at date t.*

CONDITION 4 (Stolen Thunder Condition). *Suppose that the equilibrium entails one party's making an offer of contract C at date (round) t on the equilibrium path for some $t > 1$. Then he or she would accept an offer of C by the other party in round $t - 1$.*

As Hermalin and Katz (1993) discuss, both refinements seem reasonable for bargaining games of complete information.

Given these refinements one can establish a stronger result:

PROPOSITION 2 (Hermalin and Katz, 1993). *Suppose the two parties to a contract are symmetrically informed prior to and during bargaining and that bargaining consists of alternating offers. Assume the costs of delay in achieving an agreement are due to discounting or from per-round fixed costs. If the Strict Tradeoff, Monotone Acceptance, and Stolen Thunder conditions are satisfied, and there is at least one contract that each party strictly prefers to no agreement (the status quo), then bargaining ends with their agreeing to a Pareto-efficient contract in the first round of bargaining.*

PROOF. This is Proposition 5 of Hermalin and Katz (1993). □

Propositions 1 and 2 make the case that, when society is solely concerned with the wellbeing of the parties to the contract and those parties are symmetrically informed at the time of contracting, there is no reason to believe that restricting the parties' freedom of contract will improve efficiency.

2.3. *The economic case against freedom of contract*

So far, we have focused on the economic case *for* freedom of contract. Now we review the case against. Our discussion of the Coase Theorem suggests two potential grounds on which to argue against (complete) freedom of contract: (i) actors who are not party to a contract (third parties) are affected by externalities resulting from the contract; and (ii) problems in negotiating a contract prevent the parties from writing the optimal contract.

2.3.1. *Third-party externalities*

As we saw in §2.2.2, the efficiency of markets and private contracting is contingent on there being no third-party externalities. For instance, the market equilibrium with a competitive, but heavily polluting, industry does not maximize welfare—the supply of the good in question is determined by the private costs incurred by the manufacturers rather than the social costs that account for both those private costs and the harm the pollution imposes on society. Because social costs are greater than private costs, more than the welfare-maximizing quantity gets sold.

The inefficiency of the market when externalities are present can justify restrictions on private contracts. For instance, to deal with negative externalities, society does better by restricting the freedom of buyer and seller to set price; there exists a price floor above the market-equilibrium price that forces trade to occur at the welfare-maximizing level.[13]

A strong believer in the Coase Theorem might object to this conclusion, arguing that polluters and their victims could contract to set the pollution level optimally. While that

[13] A price floor is not the only policy tool that could improve efficiency relative to the unfettered market. Other possibilities are an excise tax on the good, a pollution tax, or permitting the manufacturers some ability to cartelize their industry.

might be plausible in the context of a single polluter and a single victim (*e.g.*, noise pollution issues between neighbors), most situations of interest involve multiple polluters and millions of victims. It is difficult to imagine that significant expenditures of time and effort aren't required for a multitude of parties to reach an agreement on the terms of a contract. Moreover, as Farrell (1987b) notes, the unknown intensity of the parties' preferences typically means that any such bargaining would occur under asymmetric information. When such real-life transactions costs are accounted for, restrictions on contracts could be the more efficient means of solving the externality problem.

The transactions-cost issue is worse when the victims of the externality are not even known at the time parties enter into a contract. Aghion and Bolton (1987) nicely illustrate the problem: A monopolist (*e.g.*, a manufacturer), concerned about entry into its market, signs long-term exclusive-dealing contracts with buyers (*e.g.*, retailers). An entrant enters only if it is more efficient (has a lower marginal cost) than the incumbent monopolist. Whether such an entrant will exist is unknown *ex ante* by the incumbent monopolist and any given buyer;[14] but both know the distribution of the potential entrant's marginal cost. Because of the potential for entry, the buyer can expect to earn more surplus in the future than it currently does; as noted by Bork (1978) and R. Posner (1976), this will make the buyer reluctant to enter into an exclusive dealing contract with the seller. However, some of the surplus generated by the entry of a more efficient entrant is captured by the entrant itself. Hence, if the buyer and incumbent seller collude—that is, enter into an exclusive-dealing contract with a liquidated-damages provision that the buyer must pay the incumbent seller if it switches to the entrant—then buyer and incumbent seller can capture some of the entrant's surplus. The entrant must lower its price *vis-à-vis* the price it would have charged absent any liquidated damages provision by the amount of the liquidated damages to induce the buyer to switch to it. If the entrant lowers its price, then surplus is being transferred from it to the buyer-seller combination. The problem with such exclusive-dealing contracts is that the buyer-seller combination is like a non-discriminating monopolist,[15] it sets the liquidated damages provision too high—in the combination's desire to capture more surplus from the most efficient entrants, it deters entry from those that are only moderately more efficient. Consequently, prohibiting exclusive-dealing contracts increases expected welfare by increasing the probability of entry by a more efficient producer.

Observe that it is difficult to invoke the Coase Theorem in response to the Aghion and Bolton model. Because the entrant is unavailable at the time the incumbent seller and buyer contract, there is no possibility of their signing a three-way contract that achieves efficiency.[16]

[14] Later in their article, Aghion and Bolton relax this assumption and assume the incumbent monopolist has superior information; the implications of that assumption will be addressed later in §2.3.2.

[15] For a discussion of the welfare issues connected to a non-discriminating monopolist, see §2.3.3 *infra*.

[16] There is also a further problem insofar as the entrant's cost is its private information, so there is an asymmetry of information problem that would impede efficient contracting even if the entrant were known *ex ante*. We elaborate on this point in §2.3.2.

Following Rasmusen et al. (1991), exclusive dealing can illustrate another externality problem. Consider an incumbent monopolist who sells to N buyers. Normalize the surplus that each buyer enjoys under monopoly pricing to zero. Assume that, if there were entry, each buyer would enjoy surplus $s > 0$. Assume that entry is feasible only if an entrant can attract at least $\hat{N}$ buyers, where $1 < \hat{N} < N$. Observe, therefore, that if the incumbent can lock up at least $N - \hat{N} + 1 \equiv N^*$ buyers through exclusive-dealing contracts, the incumbent blocks entry. Consider a point in time prior to the arrival of an entrant; and consider the following offer made simultaneously by the incumbent to each of the N buyers: In exchange for signing an exclusive-dealing contract, the incumbent provides the buyer surplus $\varepsilon > 0$ (*e.g.*, the incumbent cuts its price by a small amount). The buyers respond independently and simultaneously to the incumbent. Take ε to be small enough that $N\varepsilon$ is smaller than the amount the incumbent would stand to lose should entry occur; that is, the incumbent does better paying out $N\varepsilon$ and keeping its monopoly than not paying that amount and facing competition. Note this will often entail $\varepsilon < s$.

PROPOSITION 3. *There is a Nash equilibrium in which all N buyers sign an exclusive dealing with the incumbent.*

PROOF. If a given buyer believes that the other $N - 1$ buyers will sign, then that buyer believes entry has been blocked (recall $\hat{N} > 1$). Hence, that buyer expects to get 0 if she doesn't sign and ε if she does. Because $\varepsilon > 0$, it is, thus, a best response to sign. □

The equilibrium of Proposition 3 is undesirable insofar as social welfare is reduced by the deadweight loss resulting from the preservation of monopoly pricing. Limitations on freedom of contract (*i.e.*, a prohibition on exclusive-dealing contracts) would be welfare enhancing.

As, however, Rasmusen et al. note, the equilibrium of Proposition 3 is not unique if $\varepsilon < s$. Another Nash equilibrium is for all buyers to refuse the contract—if no other buyer will sign, then signing would mean forgoing s in exchange for ε. Moreover, as Segal and Whinston (2000) point out, if one allows the buyers to form "coalitions," then the only Nash equilibrium will be the one in which all buyers refuse the contract.[17]

Unfortunately, as Segal and Whinston go on to show, there is still scope for entry-deterring exclusive-dealing contracts. For instance, it is possible that N^*s is smaller than the amount the incumbent would stand to lose should entry occur. Hence, there exists an $\eta > 0$ such that $N^*(s + \eta)$ is smaller than the amount that entry would cost the

[17] The word coalition in this context has a specific game-theoretic meaning; roughly a coalition in this context refers to a self-enforcing agreement. That is, here, an agreement to reject the incumbent's offer would be self-enforcing because it is a Nash equilibrium for all buyers to reject. For the situations below in which the incumbent offers a contract to only N^* buyers or the incumbent uses a "clever contract," then an agreement among the buyers not to accept would *not* be self-enforcing—because it is a dominant strategy for some or all buyers to accept, they would not honor their agreement with their fellow buyers; that is, those contracts are robust to the formation of "coalitions." We elaborate on this point below.

> If you, a buyer, sign and
>
> 1. if fewer than N^* buyers (including you) sign in total, then you will receive ε, but you will be *released* from the exclusive-dealing provision (*i.e.*, you can buy from the incumbent or the entrant); alternatively,
> 2. if exactly N^* buyers (including you) sign in total, then you receive $s + \varepsilon$ and you must buy from the incumbent; alternatively,
> 3. if more than N^* buyers (including you) sign in total, then you receive ε and you must buy from the incumbent.

Figure 3. A clever contract.

incumbent. Then the incumbent could offer just N^* buyers an exclusive dealing contract in which each buyer received $s + \eta$. Because $s + \eta$ is greater than what a buyer receives even if entry occurs, it is a dominant strategy for each of these N^* buyers to accept this contract.

Indeed, building on ideas in Dal Bó (in press), it is possible for the incumbent to induce all the buyers to sign an exclusive-dealing contract at a cost to itself that is arbitrarily close to zero. Consider the contract in Figure 3.

PROPOSITION 4. *If the incumbent offers the Figure 3 contract to all buyers, then it is a dominant strategy for each buyer to accept the contract. Hence, the unique Nash equilibrium is for all buyers to accept the contract, which entails the incumbent's paying a total of $N\varepsilon$ to block entry.*

PROOF. Consider a given buyer and let n be the number of other buyers it expects to sign (so $N - n - 1$ is the number of other buyers it expects not to sign). There are three cases to consider:

1. $n < N^* - 1$. If the given buyer signs, she will be released from the exclusive dealing and her total surplus will be $s + \varepsilon$. If she doesn't sign, it will be just s. Hence, she should sign.
2. $n = N^* - 1$ (*i.e.*, the given buyer is pivotal). If the given buyer signs, she will be obligated to buy from the incumbent, but she will be paid $s + \varepsilon$ in surplus. If she doesn't sign, then entry will occur, but her surplus will be just s. Hence, she should sign.
3. $n \geq N^*$. If the given buyer signs, she will be obligated to buy from the incumbent, but she will be paid ε in surplus. If she doesn't sign, then, because entry is blocked, she will get no surplus. Hence, she should sign.

Because, as shown, signing is best regardless of what the buyer thinks n is, signing is a dominant strategy. Because the given buyer was arbitrary, this holds for all buyers; that is, all buyers will sign. Because more than N^* buyers sign, the incumbent blocks entry, and does so at a cost of only $N\varepsilon$. □

Although we have presented the inefficiencies illustrated by Propositions 3 and 4 in the context of exclusive dealing, they are, in fact, examples of a broader phenomenon. Segal (1999b) considers the general issue of a single contract proposer, P, who can enter into a number of bilateral contracts with other actors, $A_1, \dots, A_N$, and derives conditions under which a set of unconstrained bilateral contracts will fail to maximize welfare due to externalities.[18] Other contexts include vertical relations (*e.g.*, exclusion of retailers or resale price maintenance), takeover battles (P is a raider and $A_1, \dots, A_N$ incumbent shareholders), debt workouts (P is "equity," which offers a debt-equity swap to the creditors, the As), and network externalities (P sells a network good and the As purchase it).

Again, a strong believer in the Coase Theorem might object to these conclusions on two grounds. First, a single grand contract among all the participants would achieve efficiency if there is no asymmetry of information. Second, a *binding* agreement among the N actors (*e.g.*, the retailers) in advance of bilateral contracting with P (*e.g.*, the incumbent monopolist) would ameliorate, if not eliminate, the problem. There are, however, a number of counter-objections:

- If N is large, then the transactions costs are likely to be so large as to make contract restrictions more efficient.
- Some of the N actors could be unknown or not yet exist (*i.e.*, similar to the problem in Aghion and Bolton).
- Contracting among the As could generate other concerns. For instance, there could be legitimate antitrust concerns if all the retailers of a good or in an area were allowed to write a contract among themselves.

2.3.2. Asymmetric information

As has been known since Akerlof's (1970) seminal work, asymmetric information between parties can result in market distortions. A number of authors (among others Aghion and Bolton, 1987; Aghion and Hermalin, 1990; Johnston, 1990; Spier, 1992; and Hermalin, 2002) have applied this idea to the issue of contract design; showing that asymmetric information between the parties at the time a contract is negotiated can lead to distortions in the resulting contract *vis-à-vis* the contract that would have been negotiated under symmetric information. Unless the equilibrium under the symmetric-information contract is, itself, second best, such distortions must imply a loss of welfare *vis-à-vis* the symmetric-information benchmark.

Whenever the parties negotiate imperfect contracts, the question arises whether there is scope for the legal system to improve matters, either by restricting the set of possible contracts *ex ante* or through appropriate court action *ex post*. Aghion and Hermalin

[18] For a general theory of bilateral contracting with externalities, see Segal and Whinston (2003).

(1990) explore the former possibility in the context of a signaling game.[19,20] Their analysis can be motivated as follows. A well-known restriction on debt contracts is that the contract cannot contain an enforceable waiver by the debtor of her right to declare bankruptcy. Indeed, many states in the US impose even further protections, allowing a bankrupt debtor to keep certain assets (*e.g.*, a car or house) under certain circumstances. Can such restrictions enhance efficiency or total welfare? To be concrete, consider an entrepreneur who needs to raise capital for a project. She knows how likely it is that her project will succeed; that is, whether it is a good project, which has a probability of failing of F_g or a bad project, which has a probability of failing of $F_b > F_g$. Using a short-hand common to information economics, call the entrepreneur with a good (alt. bad) project the good-type (alt. bad-type) entrepreneur. While the entrepreneur knows her type, a potential investor does not. His knowledge is limited to knowing that there is a probability $\theta \in (0, 1)$ that the project is good. Assume that if a project fails, it is impossible for the entrepreneur to repay the investor fully. Investing is, therefore, risky and the risk is greater investing in a bad project than in a good project. For this reason, the entrepreneur can get more generous terms from an investor, the more likely he thinks it is that her project will succeed. Consequently, the entrepreneur has an incentive to signal the investor that her project is good through the terms of the debt contract she offers the investor. Specifically, because the expected cost of a large payment to be paid if the project *fails* is greater for a bad-type entrepreneur than a good-type entrepreneur, a good-type entrepreneur can signal that she has a good project by promising a large payment to the investor should the project fail. The cost of signaling in this manner is that the entrepreneur exposes herself to considerable risk (*e.g.*, losing her house if the project fails).

Restricting the amount the entrepreneur can promise to repay in the case of failure can potentially generate Pareto superior outcomes. To see why, note that, because of

[19] Signaling games, first studied by Spence (1973), are games of asymmetric information in which the better informed party takes actions that have the potential to convey—"signal"—her information to the less well informed party. The classic example (Spence) is a worker who signals information about her ability to potential employers through the amount of education she acquires. An equilibrium of a signaling game is called *separating* if the equilibrium actions of the informed player vary with her information (*e.g.*, workers who know themselves to be more talented acquire more education than workers who know themselves to be less talented). A *pooling equilibrium* is one in which the equilibrium actions of the informed player do not vary with her information (*e.g.*, all workers get the same level of education).

[20] Section II.B.2 and Appendix C of Johnston (1990) also considers the implications of signaling on contract formation in the context of evaluating limited-liability rules. Unlike Aghion and Hermalin, Johnston is concerned with *default* rules rather than binding restrictions. However, as Ayres and Gertner (1992) argue, Johnston's emphasis on default rules undermines his arguments; Ayres and Gertner show that the choice of *default* rule is irrelevant in a world with costless contracting.

A recent paper by Anderlini et al. (2003) is another contribution to this literature. They focus on *ex post* actions by the courts, specifically whether the court should void certain contracts in some states of the world. Unlike Aghion and Hermalin, who focus on how restrictions can shift the equilibrium from an inefficient separating equilibrium to a more efficient pooling equilibrium, Anderlini et al. show how the expectation of the court's *ex post* actions creates the possibility of an efficient separating equilibrium when otherwise the equilibrium would be a less efficient pooling equilibrium.

the additional risk, an entrepreneur with a good project might prefer not to signal if silence were interpreted by the investor as meaning her project was "average" (*i.e.*, had a failure probability of $\theta F_g + (1 - \theta) F_b$). The difficulty is that, under a reasonable solution concept for the game,[21] the investor will interpret silence as evidence that the project is *bad*; and given the choice between looking good (signaling) and looking bad (not signaling), an entrepreneur with a good project will prefer to look good. If, however, signaling is restricted (*e.g.*, bankruptcy laws limit what the entrepreneur can pay in the event of failure), then not signaling is no longer informative. The investor will, therefore, treat all entrepreneurs as if they have an average project. Both types of entrepreneur are better off—an entrepreneur with a bad project now looks average, while an entrepreneur with a good project avoids the additional risks imposed by costly signaling. Because the investor is always held to his reservation utility conditional on his equilibrium beliefs, he is no worse off. Thus, restricting the possible terms of the contract would be Pareto superior.

Aghion and Hermalin formalize this argument in the entrepreneur-investor context. They go on to suggest, but not model formally, that the idea of contract restrictions eliminating "wasteful" signaling is more general than the entrepreneur-investor example. For instance, it could justify limits on penalties for breach of contract: To signal that she is very likely to be able to deliver a product on time, a good-type supplier might offer "too high" a penalty to be paid were she to be late; where "too high" means that she would be happier with a lower promised penalty if only the buyer wouldn't interpret that lower penalty as indicating she was the bad-type supplier. Barring excessive penalties would prevent the buyer from making that interpretation, which in turn would lead to a contract that both good and bad-type suppliers preferred.

While Aghion and Hermalin prove that restrictions on contracts *can* be welfare enhancing in the context of signaling models, they do not establish that restrictions will always be welfare enhancing. For some set of parameters, restrictions enhance welfare; for others, they don't. Moreover, in the latter case, the imposition of binding restrictions will reduce welfare (see Figure 5b of Aghion and Hermalin and connected discussion). Intuitively, there exist parameter values such that the separating PBE under asymmetric information replicates the equilibrium that would hold under symmetric information. In those situations, given our earlier discussion, it is not surprising that restrictions can only reduce, not enhance, welfare. Furthermore, if the contracting situation is already problematic, the fact that the informed player must signal can improve matters: Aghion and Bolton (1987), for instance, point out that the introduction of asymmetric information in their model pushes down the average liquidated damages penalty, thereby increasing the likelihood of efficient entry.

The lack of a clear normative conclusion from Aghion and Hermalin admittedly limits the practical application of their results. Eric Posner (2003a) sees this as a fatal flaw,

[21] Specifically, a Perfect Bayesian Equilibrium (PBE) satisfying the Intuitive Criterion of Cho and Kreps (1987).

supporting his overall indictment of contract theory as a guide to the law of contracts. But, as Craswell (2003) notes, there is no reason to require economic analysis to reach "all or nothing" conclusions before the analysis is useful normatively.

A signaling game is a particular kind of game of asymmetric information; in the context of contract design, it corresponds to a situation in which the informed party (*e.g.*, the entrepreneur) makes a contract offer to the uninformed party (*e.g.*, the potential investor). But there are other possible contract-offer games. An obvious alternative is for the uninformed party to make a contract offer to the informed party (a game known as *screening*). Can restrictions on contracts improve the efficiency of the outcomes of screening games?

The answer, as demonstrated by Hermalin and Katz (1993), is effectively no. The argument is as follows, let $\mathcal{C}_U$ be the set of unrestricted contracts and let $\mathcal{C}_R$ be the subset of restricted contracts (*i.e.*, $\mathcal{C}_R \subset \mathcal{C}_U$). Let C_i^* be the contract offered by the uninformed party and accepted by the informed party in the game where the relevant contract space is $\mathcal{C}_i$.[22] Let $v_P(C)$ be the uninformed party's expected utility under contract C. Because the *un*informed party cannot signal information, changing the contract space cannot change the informed party's acceptance rule. Hence, by the nature of optimization,

$$v_P\left(C_U^*\right) \geq v_P\left(C_R^*\right). \tag{2}$$

Clearly, if the inequality is strict, then restrictions on contracts cannot be Pareto improving. If expression (2) is an equality, then there is no reason for the uninformed party not to offer C_R^* if it Pareto dominates C_U^* and, hence, it is unclear why we shouldn't expect C_R^* to be offered even absent restrictions. Moreover, if the Strict Tradeoff condition (Condition 2) holds, then C_R^* cannot Pareto dominate C_U^* if expression (2) is an equality. To see this, were C_U^* dominated by C_R^*, then, by the Strict Tradeoff condition, there would exist a third contract $\hat{C} \in \mathcal{C}_U$ that the uninformed player strictly preferred to C_U^* and which the informed player would accept. But the existence of such a $\hat{C}$ contradicts the optimality of C_U^*. Therefore, C_R^* cannot Pareto dominate C_U^*. A robust conclusion, therefore, is

PROPOSITION 5. *Restrictions on contracts cannot be Pareto improving in screening situations if either (i) the uninformed party is not mean spirited; or (ii) the contract space satisfies the Strict Tradeoff condition (*Condition 2*).*

Why do signaling and screening models yield different conclusions? Externalities offer an explanation. Although the informed player in a signaling game is a single entity, one can nonetheless view her as being two (*e.g.*, the bad-type entrepreneur or the good-type entrepreneur). The fact that a bad type is potentially willing to pretend to be a good

[22] As is typical in contract theory, we can, without loss of generality, add the "refusal contract" to the set $\mathcal{C}_R$ (and, thus, to $\mathcal{C}_U$); where the refusal contract stipulates the same payoffs to the parties as would result if the informed player refused the uninformed player's offer. In other words, there is no loss of generality in assuming acceptance of some contract in equilibrium.

type forces the good type to distort the contract she offers so that she won't be mistaken for the bad type (*i.e.*, select a contract, though not ideal for her, that the bad type would *not* be willing to offer). The problem is that there is no way for the bad type to internalize this externality that her potential mimicry imposes on the good type. When, however, it is the *un*informed party who makes the contract offer, he is in a position to internalize the costs and benefits of attempting to differentiate the different types of the informed player. In essence, signaling can impose an externality, while screening cannot.

We can also consider social welfare in screening models of contract bargaining. Obviously, if restrictions are Pareto improving they also enhance welfare. However, as we've just seen, it will generally be the case that restrictions will not be Pareto improving with screening. Hence, we limit attention to the case in which expression (2) is a strict inequality. The question of whether restrictions can enhance welfare then boils down to whether the informed party's (average) gain from the restrictions exceeds $v_P(C_U^*) - v_P(C_R^*) > 0$.

To see that restrictions can enhance welfare, consider the following example. The uninformed player is a seller and the informed player is a potential buyer. The buyer's private information is his knowledge of the benefit, b, he derives from a single unit of the good being sold by the seller. Assume that b is drawn prior to contracting from a uniform distribution on [0, 1] and this fact is common knowledge. For convenience, assume the good holds no intrinsic value for the seller, so her cost is zero. While the buyer knows his benefit from purchase at the time of contracting, the seller knows only that it was drawn from the uniform distribution. The unconstrained contract offered by the seller will be a contract to sell the good at a price of $1/2$, which yields total expected welfare of $3/8$ (of this $1/4$ is expected profit for the seller and $1/8$ is the expected surplus captured by the buyer).[23] Consider the restriction that $p \leq 0$. Within this constrained space, the seller will set a price of $p = 0$. All types of buyer buy, so expected welfare is $1/2$ (all of which is captured by the buyer).[24]

Observe that the welfare loss that arises in an unrestricted world occurs because of asymmetric information. If the seller knew the buyer's valuation, then the seller would set $p = b$. All types of buyer would buy, so expected welfare would be $1/2$ (all of which would be captured by the seller). Alternatively, if the buyer did not know his valuation at the time of contracting—so the parties are symmetrically informed insofar as they both know only that $b \sim \mathsf{U}[0, 1]$—then welfare would be maximized by a contract that set $p = 1/2$ and the buyer always bought; the seller gets $1/2$ for sure and the buyer's expected surplus is zero, so total welfare is $1/2$. Note this last result is consistent with the view that asymmetries of information that arise *after* contracting are *not* justifications for restrictions; a point made more generally by Hermalin and Katz (1993)—see Propositions 1 and 2 above.

[23] At price p, the buyer will buy if $b \geq p$. The probability $b \geq p$ is $1 - p$. The seller's expected profit from a price p is, thus, $p(1 - p)$, which is maximized by $p = 1/2$. If the unit is traded, welfare is just b; hence, expected welfare is $\int_{1/2}^{1} b db = 1/2 - 1/8 = 3/8$.

[24] Expected welfare is $\int_0^1 b db = 1/2$.

In §2.2.1, we observed that the welfare criteria of Pareto efficiency and social welfare often coincide when transfers between the parties are feasible. It is worth, therefore, considering why that coincidence breaks down with asymmetric information. When there is asymmetric information, transfers are called upon to serve double duty. They continue to be a means of transferring surplus so that the welfare-maximizing allocation might be viewed as Pareto efficient by the parties. But, with asymmetric information, they are also a means of screening the different types. As our simple example illustrates, this second duty impedes transfers from doing the first optimally.

To summarize: In a signaling situation, restrictions on contracts can lead to Pareto superior outcomes and, thus, can increase total welfare. In contrast, in a screening situation, restrictions on contracts generally cannot be expected to generate Pareto improvements, although they can increase total welfare.

2.3.3. Market power

As discussed above, competitive markets can be expected to maximize welfare in the absence of externalities. When, however, one or more entities have *market power*, the market can no longer be expected to yield the social welfare-maximizing allocation. It is well known that a firm with market power (*i.e.*, that faces a downward sloping firm-specific demand) produces less than the welfare-maximizing quantity. So, at least in the standard static framework, market power reduces welfare.[25] Consequently, public policy should generally oppose contracts that promote market power over competition, such as cartel agreements. This logic extends to other contracts, such as exclusive-dealing contracts,[26] that firms might sign to maintain, establish, or extend market power.

It is important to recognize that the welfare loss that comes from market power (*i.e.*, non-discriminating or simple monopoly pricing) is an example of the adverse welfare consequences of asymmetric information in a screening model (see §2.3.2).[27] While buyers know their valuations for each unit, the monopolist does not. Hence, the monopolist sets her price both to affect the transfer of surplus from buyers to herself and to screen the buyers. As before, asking a single instrument to serve two roles leads to distortions in welfare.

[25] In a dynamic framework, one may need to consider the incentive effects of monopoly profits for innovation; that is, in some contexts, without the monopoly profits that intellectual property protection (*e.g.*, patents) afford, the innovator would lack sufficient incentive to innovate. No innovation means zero units are traded, which yields even less welfare than the monopoly outcome.

[26] Recall the discussion of Aghion and Bolton (1987) and Rasmusen et al. (1991) in §2.3.1.

[27] It follows from the Coase Theorems (Theorems 1–3) or Propositions 1 and 2, as appropriate, that bargaining power should *generally* be irrelevant in a standard welfare analysis *absent* asymmetries of information. This is not to say that there couldn't be other social reasons for concern about inequities in bargaining power, such as distributional concerns (consider, *e.g.*, Kennedy, 1982). As E. Posner (2003a) notes, courts have been known to declare contracts unconscionable because of unequal bargaining power. See also §2.4.1 and §2.5.2 *infra*.

The fact that the monopoly-pricing problem is a screening problem also implies that if the monopolist knew each buyer's valuation schedule for all the units he could conceivably wish to purchase, then it could achieve the welfare-maximizing allocation. That is, as is well known (see, *e.g.*, Tirole, 1988, §3.1), a perfect price-discriminating monopolist maximizes social welfare.[28] Hence, the welfare loss typically seen with monopoly stems not from market power alone, but from the combination of market power and asymmetric information.

Price discrimination in the real world is imperfect. Depending on the form of discrimination, the structure of demand, and other circumstances, allowing a monopolist to engage in price discrimination versus simple monopoly pricing can enhance or diminish welfare (see Chapter 3 of Tirole, 1988, for a review of the welfare consequences of imperfect price discrimination). Consequently, it is difficult to assess, at a general level, what policy should be towards the enforceability of contractual terms that facilitate price discrimination.

For example, it is not obvious from a welfare perspective whether airlines should be free to issue tickets that contain restrictions (*e.g.*, an obligation to stay a Saturday night before returning to one's point of origin).[29] Such matters need to be studied on a case-by-case basis.

Note too that market power is connected to bargaining power. Indeed, the screening problem associated with a non-discriminating monopolist can be interpreted as stemming, in part, from the seller having all the bargaining power, so that she gets to make a take-it-or-leave-it offer to buyers. The welfare loss from monopoly would disappear if it were the buyers who could make take-it-or-leave-it offers (assuming they knew the seller's marginal-cost schedule).

2.3.4. Capacity and bounded rationality

The parties in traditional law & economics analyses are presumed to be sophisticated and to possess the requisite capacity. Consequently, any contract into which they voluntarily enter is, in rational expectation, superior for them to their no-contract (*status quo*) position. That is, each party correctly estimates that their expected utility from contracting exceeds that from not.

Observe that the rationality being assumed has two components. First, neither party would enter into an agreement that he or she thought would make him or her worse off,

[28] Of course, because a perfect-discriminating monopolist captures 100% of the surplus, there could be distributional or equity grounds for this outcome not to be favored.

[29] Saturday night restrictions are a form of second-degree price discrimination whereby an airline can screen business travelers (those with a high value of flying and a high cost of staying over Saturday) from non-business travelers (those with lower values of flying and lower costs of staying over Saturday). If banning such restrictions caused the airlines to price so that far fewer non-business travelers flew, then the ban would almost surely be welfare reducing. If, however, the ban led the airlines to price so as to keep non-business travelers flying, then eliminating the distortionary effects of the Saturday-night restriction would almost surely be welfare enhancing.

in expectation, than not contracting. Second, each party is forming these expectations in an objectively correct manner. For instance, if you respond to some get-rich-quick spam email, you presumably expect to enrich yourself, but such expectations are not rational; that is, you are rational in the first sense, but not the second.

It is difficult to argue that people aren't rational in the first sense (putting aside certain pathologies such as compulsive self-destructive behavior), but there has long been some unease with the assumption that people are rational in the second sense (see, *e.g.*, Simon, 1972, or Rubinstein, 1998, for a discussion). More recently, a movement has arisen within law & economics generally (see, *e.g.*, Jolls et al., 1998 and Korobkin and Ulen, 2000) that also questions the assumption of rationality in the second sense. This movement has been labeled "behavioral law and economics," and it resembles related work in economics in its emphasis on cognitive errors, framing effects, time-inconsistent patterns of discounting, and similar phenomena. Within the formal modeling of contracts, however, assumptions of *bounded rationality* have been largely limited to the issue of explaining incomplete contracts (or justifying assuming that contracts are incomplete). We take up this use of bounded rationality in §4.2.1 below.

While, to the best of our knowledge, capacity and sophistication have not been formally modeled in the context of contract theory, such issues have received attention by law & economics scholars in law reviews (see, *e.g.*, Eisenberg, 1995; Korobkin and Ulen, 2000; and Bar-Gill, 2005).[30] Korobkin and Ulen, for instance, argue that mandated contractual terms can be justified when one side lacks sophistication. They take as an example the former practice of insurance companies to cover only one day of hospital stay for maternity and the legislative reaction that required these companies to cover longer stays. They argue that legislative action was necessary because "the possible permutations of coverages that could, in principle, be provided by a given health insurance policy are numerous and . . . can overwhelm even sophisticated consumers . . . [and their] agents . . . who might be making purchasing decisions" (Korobkin and Ulen, p. 1082).

While it is difficult to argue against the proposition that the majority of consumers do not fully understand the provisions of their insurance contracts,[31] the example of maternity benefits seems a poor one from which to argue for intervention on the basis of inattention. The legislative reaction to this issue, and the strong public outcry that prompted it, suggest that the limited coverage insurers had previously been offering was not due to any consumer inattention. A more likely explanation was insufficient willingness to pay.[32] Put a bit differently, given the apparent awareness of the market to

[30] Although, as we discuss *infra*, some formal models (*e.g.*, Katz, 1990b and Rasmusen, 2001) touch on related issues.

[31] See, for example, Eisenberg's (1995) discussion (p. 242) of *Gerhardt v. Continental Insurance Cos.* (48 N.J. 291, 295 A.2d 328), in which he quotes the justices of the New Jersey Supreme Court expressing their difficulties in understanding the terms of the insurance contract in question.

[32] Korobkin and Ulen cite statistics (their footnote 111) that 90% of insured Americans get their health insurance through their employer. This fact, however, is potentially problematic for their argument. The employers

how much maternity stay was being covered, one would have expected the "unregulated market" to provide the desired level of maternity benefits if maternity benefits had been important to those who actually decided about policies.

A more plausible variant of the Korobkin and Ulen idea is developed by Eisenberg (1995). Although Eisenberg presents his argument verbally rather than mathematically, it is helpful to add some degree of formalism.[33] Suppose there are two possible future states, a rare one, which occurs with probability r; and a more likely one, which occurs with probability $1 - r$. Eisenberg is interested in the case where r is small, but not zero. Consider contracting between two parties, P and A. For convenience, assume that the bargaining between them always results in a contract that yields monetary benefits B_i (gross of direct transfers) to party i in the more expected state (*i.e.*, the state that occurs with probability $1 - r$). Suppose that the contract can contain a stipulation, $s \in \mathcal{S}$, that governs what happens in the rare state. Let $b_i(s)$ denote party i's monetary benefit gross of direct transfers in the rare state under stipulation s. Assume that the stipulation that maximizes P's benefit in the rare state, $\hat{s}$, is not the stipulation that would maximize the sum of the parties' benefits. Let that stipulation be s^*. Mathematically,

$$\begin{aligned} b_P(\hat{s}) &\geq b_P(s) \text{ for all } s \in \mathcal{S}, \\ b_P(\hat{s}) &> b_P(s^*), \\ b_A(s^*) + b_P(s^*) &\geq b_A(s) + b_P(s) \text{ for all } s \in \mathcal{S}, \quad \text{and} \\ b_A(s^*) + b_P(s^*) &> b_A(\hat{s}) + b_P(\hat{s}). \end{aligned} \tag{3}$$

Observe that there must be $s \in \mathcal{S}$, including s^*, that A prefers to $\hat{s}$. Define $\Delta_i = b_i(s^*) - b_i(\hat{s})$. By construction,

$$\Delta_A > \Delta_A + \Delta_P > 0 > \Delta_P.$$

Eisenberg is interested, in part, in situations in which P proposes the stipulation for the rare state and it costs A some amount, k, to evaluate or understand the stipulation that P has proposed. One can view k as the cost of consulting an attorney or the cost of time spent trying to understand the small print. If A is particularly unsophisticated, one might even set $k = \infty$ to reflect the idea that A is simply incapable of understanding what P has proposed. Suppose that if A does understand the stipulation, then P and A would bargain to some term that A found acceptable. In keeping with our earlier analysis of bargaining under symmetric information, we may as well assume that they

could be quite sophisticated and simply shopping for a bargain in insurance. That is, the problem is due to agency (the employers obtaining insurance on their workers behalf) rather than bounded rationality. Korobkin and Ulen might argue that workers suffer the same cognitive limitations bargaining with their employers; but, depending on worker-employer bargaining, the relative political power of workers versus employers and insurers, and the incidence of the cost of increased medical coverage, it is nevertheless possible that everything can be explained within the rational-actor paradigm.

[33] Katz (1990b) and Rasmusen (2001) are two articles that also model "reading costs" formally.

would then agree on s^*. A critical assumption, however, is that

$$k > r\Delta_A; \tag{4}$$

in other words, the expected gain that A stands to reap from understanding the stipulation for what should happen in the rare state is smaller than the cost of obtaining that understanding (we will see in a moment that, when A doesn't understand the stipulation in the rare state, P should propose $\hat{s}$). Observe that expression (4) holds if k is large or if r or Δ_A are small; Eisenberg can be read as being primarily concerned about the situation in which r is small—we will return to this point later.

To close the model, suppose that the bargaining process is such that P can be seen as proposing a contract subject to A's acceptance, where A accepts only if his expected net benefit at least meets a reservation level β_A; that is,

$$(1 - r)B_A + r\tilde{b}_A - p \geq \beta_A, \tag{5}$$

where $\tilde{b}_A$ is the benefit A *expects* in the rare state and p is a transfer to P (if $p < 0$, then the transfer is from P to A). P's utility is

$$(1 - r)B_P + r\tilde{b}_P + p,$$

where $\tilde{b}_P$ is what she gets in the rare state.

Suppose that A will not expend the k necessary to understand the stipulation for the rare state; then it is irrelevant for A's acceptance rule (5) what stipulation P *actually* makes because A won't understand it and, thus, his expectation, $\tilde{b}_A$, cannot depend on it. Given this, it is clearly a best response for P to choose $\hat{s}$.[34] In equilibrium, A must form his expectations correctly,[35] so $\tilde{b}_A = b_A(\hat{s})$. Observe, given (4), it is indeed a best response for A not to expend k on understanding even if A anticipates P is proposing $\hat{s}$. In equilibrium, total welfare will, thus, be

$$\widehat{W} \equiv (1 - r)(B_A + B_P) + r\big(b_A(\hat{s}) + b_P(\hat{s})\big).$$

From expression (3), it follows that total welfare is *not* being maximized.

If we suppose that P has sufficient bargaining power so that expression (5) is binding, then P's equilibrium utility is

$$(1 - r)(B_A + B_P) + r\big(b_A(\hat{s}) + b_P(\hat{s})\big) - \beta_A,$$

from which we see that it is P who bears the full cost of this loss in welfare. Normally when a party bears the full cost, that party would take steps to eliminate the efficiency

[34] One might wonder why P bothers to worry about what to stipulate if the state really is rare. One answer, suggested by Eisenberg, is that P deals with many As (*e.g.*, the contract is a form contract that P uses with many customers); hence, even though the expected gain to P from stipulating $\hat{s}$ over s^* is small for any one transaction, in aggregate it is large enough to induce P to invest in choosing the best stipulation for her.

[35] This imposition of rationality on A might seem at odds with the bounded rationality approach of Eisenberg. We will discuss this point later.

loss. However, here, P can't—there is no credible way for her to commit to A that she has stipulated s^*; that is, any promise of s^* is cheap talk because P knows A will not understand if P substitutes the stipulation that favors her.

There are some ironies in this result. First, while both Korobkin and Ulen and Eisenberg appear to suggest that it will be the *un*sophisticated party who suffers in these situations, the truth could well be that it is the sophisticated party who suffers. Second, even though, as noted by Schwartz and Wilde (1979), the sophisticated party has a motive to internalize the value of the superior contract term, there is no way for her to do so.[36]

Conversely, suppose that, due to competition among the Ps, it is the As who capture the surplus; that is,

$$(1-r)B_P + rb_P(s) + p = K_P,$$

where K_P is P's reservation utility (cost). Observe this expression can be rearranged as

$$p(s) \equiv K_P - (1-r)B_P - rb_P(s). \tag{6}$$

From expression (6), it follows that the price a P charges if the term is $\hat{s}$ is lower than if it is s^*; that is, $p(\hat{s}) < p(s^*)$. In this environment, we can assess Schwartz and Wilde's claim that if some of the As have a low enough cost of evaluating the stipulation for the rare state, then competition will lead the Ps to stipulate s^* rather than $\hat{s}$. Suppose there are two kinds of As, those with low evaluation costs, k_ℓ, and those with high evaluation costs, k_h. For the moment, set $k_\ell = 0$, so the ℓ-type As always evaluate. For the h-type As, k_h satisfies expression (4). Suppose the presence of the ℓ types led all Ps to stipulate s^*. Competition among the Ps would then require them each to charge $p(s^*)$ and earn zero profits. Consider a deviation from this candidate equilibrium in which a P stipulated $\hat{s}$. Although she would lose all business from ℓ types, she would, at least in the short term, keep the h types. Moreover, at a price of $p(s^*)$, she would realize an expected gain of $-r\Delta_P > 0$ from a contract with an h type. Hence, deviating from the candidate equilibrium—stipulating $\hat{s}$ rather than s^*—is profitable for that P, which means the candidate equilibrium isn't an equilibrium at all. In other words, there is no equilibrium in which all Ps stipulate s^* for sure.[37]

In fact, if we set $k_\ell > 0$, we can see a second reason why an s^*-only equilibrium cannot arise. Define $\tilde{\Delta}_A = b_A(s^*) - \tilde{b}_A$; that is, $\tilde{\Delta}_A$ is the expected gain for A that he expects from negotiating the contract from the expected stipulation in the rare state to

[36] In Katz (1990b), the sophisticated party can take actions costly to it that lower the unsophisticated party's cost of understanding the contract terms. On the other hand, if those costs are incurred on a per-transaction basis (*e.g.*, they represent the expense of a person explaining terms to a potential buyer), then such actions could well fail to be cost effective, especially if the terms in question pertain to a rare event.

[37] What *is* the equilibrium in this game depends on what additional assumptions we make. If, for instance, we assumed that ℓ types could, in negotiations, change the contract from $\hat{s}$ to s^* in exchange for an increase in price from $p(\hat{s})$ to $p(s^*)$, then an equilibrium can be constructed in which all Ps offer contracts at $p(\hat{s})$, but end up negotiating different terms with the ℓ-type As.

s^*. Observe, even for an ℓ type, it is only worth evaluating the stipulation if

$$k_\ell \le r\tilde{\Delta}_A.$$

Suppose all Ps were expected to stipulate s^*; then, $\tilde{\Delta}_A = 0$. Hence, no types of A would bother to evaluate the stipulation. But, as we've seen, if A won't evaluate, then P should always stipulate $\hat{s}$. The conclusion is, thus, that it is impossible for A ever to be certain that P has stipulated s^* unless A checks (we might dub this the "Eisenberg uncertainty principle").[38]

Although this model is somewhat rudimentary, it does demonstrate that the problem identified by Eisenberg and others is unlikely to be eliminated by private action. This suggests that intervention could be beneficial. Such intervention could be *ex ante*, *i.e.*, the law could disallow any stipulation but s^*; or the intervention could be *ex post*, *i.e.*, the courts could grant A damages should it be shown that the stipulated term was other than s^*.

As we observed earlier, a critical condition is (4), which makes it rational for A not to seek to understand the terms stipulated for the rare state. If condition (4) fails, as it will if the relevant terms (r or Δ_A) are large, then the argument for intervention collapses.[39] While either r or Δ_A small enough is sufficient for condition (4) to hold, the situation in which it holds because Δ_A is small could be relatively uninteresting—presumably A will need to take some action to enforce restrictions on s (*e.g.*, sue to have an illegal contract voided or to recover damages); but if Δ_A is small, then A's motive to do so could be small as well. That is, it could be difficult to implement intervention in cases in which Δ_A is small. The situations in which r is small are, as Eisenberg suggests, the more important ones. The *ex post* injustice could be large, but the motive of a single A to address it *ex ante* could be too small to make feasible a private solution.

Admittedly, we have imbued A in our formal model with rather more rationality than Eisenberg intended. Like others (*e.g.*, Jolls et al., 1998; Korobkin and Ulen, 2000; Bar-Gill, 2005), he observes that research in psychology and behavioral economics suggests that most people are poor at estimating probabilities correctly, especially probabilities of small events.[40] To the extent that they systematically underestimate the probability of a rare event (*e.g.*, likelihood of an insurance claim), they will use too small an r in deciding whether to invest in understanding what the contract stipulates for rare states. While the addition of such cognitive biases have the potential to enrich the analysis, our simple model demonstrates that they are not essential to obtain the main conclusion.

Katz (1990b), among others, notes there is a relation between the analysis in this subsection and Akerlof's lemons model. Although, here, we have a problem of moral hazard—the contract proposer *chooses* her "type"—whereas Akerlof is concerned with

[38] Katz's (1990b) Lemma 2 makes the same point.

[39] Although, even when (4) fails, there could still be cause to be concerned due to the Eisenberg Uncertainty Principle.

[40] See Rabin (1998) for a survey.

a problem of adverse selection—the informed party's type is *exogenously* determined—it is true in both that the uninformed party expects adverse terms of trade. A major difference, however, is that here all valuable exchanges take place, but just not under the optimal contract. In the lemons model, in contrast, buyer suspicion about quality can push the price below the level at which high-quality sellers are willing to sell, leading to inefficient exit of the high-quality sellers from the market. Admittedly, if the terms in question were important enough to the uninformed party, but yet not important enough to incur the reading costs, then it is theoretically possible that no trade could occur because of the uninformed's suspicions about the adverse nature of the contract. These issues have been discussed informally in the legal literature as well (see, for example, Kennedy, 1982; Craswell, 1993, 2000; and Eisenberg, 1995).

2.4. Other arguments for regulating private contracts

2.4.1. Distributional fairness

The Second Theorem of Welfare Economics holds that given convex production and utility functions, complete markets, and costless redistribution, any Pareto efficient allocation can be implemented as the general equilibrium of a competitive economy (see, *e.g.*, §6.4 of Debreu, 1959). In the real world, however, these necessary assumptions are frequently violated, so that the same policy instruments must be used to promote both efficiency and distributional justice, with the two goals being traded off in a second-best fashion (Okun, 1975). Accordingly, much economic regulation has in practice some distributional component (R. Posner, 1971, and Polinsky, 1974).

In the field of contracts, however, as opposed to other bodies of law such as tort or property, it is relatively difficult to use private law rules as instruments of distribution. This is because contractual obligations are in general voluntarily undertaken, so that changes in legal rules will be accompanied by changes in the parties' reservation prices for exchange, tending to shift the incidence of a regulation to its intended beneficiary (*e.g.*, Buchanan, 1970). But restrictions on price will redistribute between marginal and inframarginal market actors, and in cases of market power, can redistribute between the two sides of the market (as can readily be seen from the example of a monopolist subject to a price ceiling set between the monopoly price and marginal cost). And restrictions on non-price terms can similarly redistribute between marginal and inframarginal actors if they differ in the value they attach to such terms (Spence, 1975; Craswell, 1991).

Similarly, policies that internalize externalities have distributional effects as well, and from a political economy viewpoint such effects—because they are larger—are probably more important than any efficiency gains in explaining the pattern of regulation. For example, in situations of adverse selection (the most important externality arising in the contractual setting), policies that force information revelation or limit entry or exit from a market can cause redistribution across informational types.

2.4.2. Liberty and autonomy

In the context of contract law, the liberal perspective tends to support arguments for contractual freedom, although on libertarian rather than economic grounds. As indicated in §1.2.3, however, liberal and economic theories of contract can diverge in cases of market failure or transaction costs. A canonical illustration is provided by the phenomenon of standard-form contracts. Standardized contracts are an inevitable by-product of a mass production economy, but many legal commentators (*e.g.*, Kessler, 1943 and Rakoff, 1983) have regarded them with distrust on the grounds that most persons presented with standard forms quite reasonably do not bother to familiarize themselves with the specific contents, relying instead on the drafter's reputation and on the knowledge that other contracting parties regularly do business on like terms. Such commentators take the position that these facts vitiate the non-drafting party's consent and justify state regulation of fine-print terms, although other writers, more influenced by an economic or commercial perspective (*e.g.*, Llewellyn, 1960, pp. 362–371) are willing to read implied consent into the overall situation. This issue is discussed at greater length in §3 and §4 below, where we take up the problems of contract formation and interpretation.

2.4.3. Inalienability and commodification

Finally, both popular morality and legal institutions commonly limit transactions dealing with matters thought to be fundamental to citizenship or personal identity; common examples include prohibitions on slavery, sexual prostitution, and the transfer of political rights such as suffrage or military service. The entitlements subject to such restriction are often described in political terms as inalienable, a concept not easily incorporated into economic accounts of exchange. Sometimes, as in the case of sexuality or body parts, restrictions are imposed on market exchange but not on other transfers: a situation that Radin (1987) labels "market-inalienability."

In some cases, restrictions on alienability can be justified in terms of market failure such as asymmetric information (*e.g.*, the use of in-kind benefits to screen out those not in need, as in Blackorby and Donaldson, 1988) or externality (when Pigouvian taxes are administratively infeasible). But many such restrictions are better explained by the idea that exchange of the relevant entitlement injures some fundamental interest of the restricted agent not captured by his utility function, or some social interest that cannot be translated into material or pecuniary terms. In economics, this idea finds historical roots in the classical Marxian idea of commodification, or the change in the social meaning of a good that occurs when it becomes the subject of economic exchange.

The concept of commodification can be translated into neoclassical economic terms by interpreting it as sort of a cultural externality (*e.g.*, when sexual prostitution is said to diminish the quality of other people's relationships) or as a lexicographical preference ordering on the social welfare function (so that equality in the distribution of certain goods trumps other allocative or distributional considerations, as in Tobin, 1970). But this interpretation does not do justice to the sense in which those who advocate the concept are concerned with the social construction of perceptions and preferences. Accordingly, the concept is best explained (see, *e.g.*, Kelman, 1981) as relating to potential changes in the content of individual utility functions or of the cumulative social welfare function.

2.5. Legal doctrines regulating freedom of contract

To a considerable extent, the actual legal doctrines restricting freedom of contract can be understood in economic terms, in that most doctrinal restrictions roughly correspond to situations of market failure or high transaction costs of the sort discussed in the previous subsections. This correspondence is not exact, however; and a survey of the main restrictions will illustrate the roughness of the relationship.

2.5.1. Formal requirements for contracting

The law imposes a number of formal requirements that must be satisfied in order for contractual agreements to be enforceable in court. In general, as discussed in §2.1.1 above, failure to satisfy these requirements results merely in denial of public enforcement rather than any negative sanction. Under US law, for instance, a contract for the sale of goods with market value above $500 (a threshold that has not been adjusted for fifty years) is not ordinarily enforceable in the absence of a writing.[41] Nonetheless, there are no penalties for entering into an oral contract of this sort and in fact such bargains are regularly concluded and performed without incident.

In many other situations, though, denial of public enforcement operates as a binding sanction, especially in one-shot deals where the potential gain or loss from breach of contract is high compared to the value of future business relations. For instance, few businesses would use oral contracts to govern the production and sale of custom-made machinery; and fewer lawyers would advise their use. In such cases, formal requirements can have important regulatory consequences; and their costs can deter exchanges from taking place (as when a business decides not to go public because of the burden of complying with the SEC's disclosure and auditing requirements).

Most legal formalities, like many default terms of the sort discussed above in §2.1.3 are justified, at least in part, by the need for coordination or by administrative needs of the legal system. The rules of offer and acceptance, for instance, discussed below in §3, are typically interpreted as social conventions that serve to help contracting parties ensure that they attach similar meaning to their words and actions and that this meaning will be understood by third parties interested in the agreement. Similarly, the law encourages parties to structure their arrangements and to create and present evidentiary records in a way that lowers the *ex post* costs of contract enforcement—both because of the economies of scale involved in setting up any particular administrative regime,

[41] UCC 2-201.

and because it is judged infeasible or politically undesirable to tailor court fees to the individual costs of dispute resolution.

Many formalities, however, also serve regulatory purposes. For example, the requirement that litigation documents be submitted on standard size paper (a formality imposed by virtually all court systems) serves purely administrative purposes, but the requirement that a minor's application for a marriage license be signed by a parent or guardian is also intended to help prevent the minor from making a hasty and irrevocable decision. And some formalities, such as the doctrinal rule that silence in the face of an offer does not ordinarily constitute acceptance, serve both coordinating and regulatory functions.[42] Such a rule indeed comports with standard interpretive conventions, but it also serves to discourage parties from sending purely opportunistic offers in the hopes that the recipient will somehow overlook them.

When viewed as restrictions on contractual freedom, formalities are plausibly justifiable on at least three of the theories of market failure discussed in §2.3 above. First, formalities can operate to correct informational problems. Parties entering into contracts may not always realize that they are doing so, especially if they are amateurs or newcomers to the relevant commercial community; similarly, they may not understand all of the specific obligations entailed by contracting. Formalities that warn such parties against assuming unintended or unwanted obligations can thus prevent inefficient exchanges as well as undesired distributional transfers from the uninformed to the informed. (On the other hand, as many anti-formalist critics have argued, hidden formalities can also mislead naïve agents into thinking they have entered into binding legal obligations when in reality they have not.)

Second, formalities may provide an effective response to bounded rationality if their presence triggers some cognitive or institutional process that operates as a safeguard against the specific dysfunctional behavior at issue. For example, the federal regulation requiring a "cooling-off" period before completion of a consumer door-to-door sale allows time for additional reflection at a remove from any pressure imposed by the salesperson; and in an organizational setting, writing requirements allow principals more easily to know when an obligation has been undertaken and to monitor their agents for making bad bargains.

Third, formalities can help to correct various negative externalities, especially those that are moderate in magnitude or related to informational asymmetry, by attaching an additional cost to the externality-producing behavior. In markets affected by adverse selection, for instance, formalities that make it harder to disclaim warranties or opt out of medical coverage may keep some high-quality agents from exiting the market, supporting a higher quantity of exchange. Additionally, uncertainty regarding the existence and enforceability of contracts may in some settings adversely affect third parties whose economic fortunes are linked with the contracting agents. Such externalities may explain the traditional common-law rule requiring a writing for contracts for the sale of

[42] See Restatement (Second) of Contracts 69 (1979). In exceptional circumstances, however, such an inference is justified, as when there has been a course of dealing that leads the offeror to expect a response.

land or in consideration of marriage—both transactions with potentially significant implications for other members of the contracting parties' family and community. Finally, formalities can help deter rent-seeking by opportunistic agents hoping to take advantage of the informational and behavioral problems discussed just above.

To illustrate these possibilities, we briefly discuss two of the more important formalities imposed by the common law of contracts: the doctrine of consideration, and the statutory requirement that certain contracts be executed in the form of a writing.

The doctrine of consideration. In the US and in other legal regimes that descend from the English common law, contractual promises are not enforceable at law unless supported by what lawyers call consideration. The precise meaning of this concept has long been debated, but its overall import is to require the contract to relate to an exchange that is understood as such by both parties. A promise to make a future gift is not enforceable under the law of contracts, for instance, although it may be possible to achieve a similar result by using some other legal device such as an equitable trust. While the consideration doctrine has at times been viewed as a substantive limit on contractual freedom, modern commentators (following Holmes, 1897, and Fuller, 1941) view it largely as matter of form, in that it can often (though not always) be avoided by the use of some other formal device such as a special writing, the transfer of a nominal sum, or delivery of a symbolic token. As such, it raises the relative cost of entering into those contracts to which it applies.

Although its grounding in exchange would seem to suggest a close connection between the consideration doctrine and the promotion of economic welfare, the doctrine would seem to diverge from simple efficiency in at least two respects. First, many non-exchange promises also enhance economic welfare. A donor's commitment to make a gift, for example, enables the beneficiary to engage in specific anticipatory investment, thus lowering the donor's cost of providing the beneficiary with any given level of utility (R. Posner, 1977 and Shavell, 1991). Second, the lawyer's understanding of what counts as an exchange is narrower in practice than an economist's would be. In traditional common law, for instance, a promise to guarantee the debt of another person was not enforceable unless the party seeking to enforce could show that the promisor received a specific promise or benefit in exchange for the guaranty; it did not suffice to observe that commercial parties do not typically make such guaranties unless they stand to gain from the extension of credit and are hoping to induce it. Similarly, certain open-ended promises, such as promises to buy or sell goods in a quantity to be specified by the promisee, were traditionally deemed unsupported by consideration, on the theory that the promisee retained discretion to fix terms that would eliminate all benefits received by the promisor.

The doctrine of consideration has long been controversial; and civil law systems, such as those in continental Europe, do not employ it. In its application to donative promises, the doctrine is probably best justified as a response either to bounded rationality or to *ex ante* rent seeking by potential beneficiaries. It is at least plausible that impulsive promises are a greater problem when motivated by generosity rather than self-interest, and requiring more elaborate formalities before such promises become enforceable may be justified in order to protect the interests of the donor or of other potential beneficiaries of his largesse who might fare better upon more thorough deliberation. Similarly, there is an obvious incentive for overreaching by potential beneficiaries, and requiring an additional formal protection such as an equitable trust, which imposes fiduciary duties on the trustee and in practice requires the participation of a legal professional to assure its effectiveness, probably serves to police such opportunism better than a mere writing requirement would.

Writing requirements. While written documents are typically regarded as better guides to the intention of the parties than oral testimony or circumstantial evidence, the general common-law rule is that most contractual promises are enforceable without a writing, if they can be proven to a court's satisfaction. Certain categories of contracts, however, must be evidenced in writing to be legally enforceable. In common-law systems, such contracts include contracts for the conveyance of interests in real property, contracts in consideration of a party's marriage, third-party guaranty contracts, contracts in which an executor agrees to pay the debts of an estate, contracts that cannot be completed within a year, and, as indicated above, contracts for the sale of goods above a monetary threshold (in the US, $500). These various writing requirements are collectively referred to as the "Statute of Frauds" because they derive originally from a parliamentary statute of 1677 that was justified by its proponents as a safeguard against false claims of contractual agreement. Similar writing requirements have been extended by statute to a variety of other transactions, including contracts that designate property as collateral for a secured loan and contracts for the sale of any personal property with value over $5000; and in some settings, such as option and guaranty contracts, a written promise suffices as a substitute for consideration.

These various requirements surely serve the purposes of *ex post* efficiency in the enforcement setting; for example, courts might well wish to screen out contractual disputes where the key evidence is not just oral but stale (the usual rationale for the one-year provision of the Statute of Frauds). Given that the interpretation of oral evidence is typically more uncertain, allowing high-value contracts to be alleged on purely oral evidence is particularly costly in terms of incentives for *ex post* rent seeking. And given the considerable private incentives for contracting parties to memorialize their agreement in writing even apart from the prospects of a dispute, the absence of a writing justifies, on Bayesian grounds, the inference that further proceedings are unlikely to be of much value.

Most of these requirements, however, also correspond to some standard type of market failure such as externality or bounded rationality. Guaranty contracts, secured lending, and contracts for the conveyance of land, for instance, all implicate the interests of third parties, and in former times, so did contracts in consideration of marriage. Requiring such transactions to be evidenced in writing reduces information costs for third parties trying to determine the status of assets in which they may also have a claim. Similarly, the expected costs of bounded rationality are especially great in high-value contracts, contracts likely to be entered into under circumstances of emotional stress, and contracts that reach far into the future. In such cases, the extra transaction costs of a formal writing may be justified because they deter impulsive or myopic decisions.

2.5.2. Substantive limitations on contracting

In addition to the practical restrictions imposed by formal requirements for contracting, some agreements are simply unenforceable as a matter of substance. Such limits on enforceability could serve to strike the entire contract, providing an affirmative defense to any liability for breach; or they could strike certain terms of the contract only, leaving the rest of the contract fully enforceable but with the offending terms replaced by default terms, or by terms the court deems fairer (see Craswell, 1993, for a discussion). The major defenses we consider here include fraud, lack of capacity, mistake, duress, undue influence, unconscionability, and offense to public policy.

Fraud and unilateral mistake. All legal systems refuse to enforce contracts that are based on sufficiently incorrect or asymmetric information, at least in cases where the uninformed party is unaware of his informational disadvantage. Enforcement is especially disfavored when one party to the exchange has caused the other to become misinformed, for example by misrepresenting material facts relating to a proposed exchange. Under the common law, this disfavor is reflected in the defense of fraud.

The fraud defense is justified on efficiency grounds for two reasons. First, it deters inefficient exchanges that would not have taken place but for the fraud; and second, in cases where fraud takes the form of undertaking efforts to deceive others, it discourages rent-seeking. Determining what statements count as fraudulent, however, is not always easy. Vague or ambiguous statement may raise interpretation issues of the sort discussed *infra* in §4; and even statements that are literally true may be interpreted as making an implied fraudulent representation. For economic discussions of this issue, see Ayres and Klass (2005) and Craswell (2006).

Additionally, the defense is not limited to cases of affirmative deception: it can also be asserted when one party withholds critical information that the other reasonably expects to be disclosed. Such fraud by omission is controversial from a legal perspective because of the difficulty of determining when there is a duty to disclose, and from an economic perspective because the duty to disclose information will affect incentives to acquire it.

Similarly, the doctrine of unilateral mistake allows a party to escape a bargain if his assent was based on a significant misapprehension that the other party could have easily corrected (as when a contractor underbids a job because of a calculation error that is evident from the large discrepancy between the mistaken bid and all others received). Here the law takes the position that the small effort required to correct the mistake is justified by the dislocation imposed on the mistaken party and the significant chance that the bargain is inefficient.

In order for doctrines like fraud or unilateral mistake to be welfare-improving in practice, however, courts must have a reliable way to distinguish between asymmetric information on the one hand, and conscious risk-taking in the setting of incomplete but symmetric information on the other, else too many *ex ante* efficient transactions will be avoided. In this area, accordingly, there is potential for divergence between what the law provides and what would be most efficient from a second-best perspective. In general, the courts have been sensitive to such concerns and have limited the scope of the doctrines to fairly significant imbalances of information (and in the case of fraud, have imposed special requirements of proof and pleading).

Incapacity. When one of the parties to a contract is incapable of weighing costs and benefits, there is no longer any basis to presume that the contract is efficient. Similarly, the law declines to enforce contracts entered into by children, intoxicated persons, and the mentally disabled; this result also deters rent seeking by those who would take advantage of the incompetent. In contrast to the situation of fraud and mistake, however, there are circumstances in which an incompetent person will be held to his bargain. One such circumstance is when the contract is for the sale of "necessities" (*e.g.*, clothing or shelter that are judged to be in the objective interest of the buyer). Another is when the competent party could not have known that the counterparty lacked capacity, although many courts will still supervise the bargain in such cases to ensure its substantive soundness.

The different treatment of the capacity and fraud defenses is justified by the differing costs of overcoming the specific transactional failure at issue. In the case of fraud, the failure can be overcome at no cost by abstaining from misrepresentation, and in the case of unilateral mistake, the failure can be overcome at low cost by correcting the mistake. In cases of incapacity, however, it may not be possible to correct the problem without losing the transaction entirely, so the law allows plainly efficient exchanges (or in the case whether the incapacity is hidden, exchanges where it is plain that no rent-seeking took place).

Mutual mistake. In some cases, the law refuses to enforce bargains in which the parties' *ex ante* beliefs were sufficiently different from the state of the world, even when those beliefs are the same and when neither party could easily have corrected the error. For example, in the celebrated case of *Sherwood v. Walker*, the court allowed the seller of a pregnant cow to void the deal on the grounds that, at the time of the bargain, both parties mistakenly thought that the cow was barren and set a price that corresponded to its slaughterhouse value. This doctrine, known as the defense of mutual mistake, is subject to much confusion and there is no canonical economic explanation of it. The legal confusion results because it is difficult in practice to know whether the parties were truly mistaken or just engaged in a hedging transaction. From the economic viewpoint, Posner and Rosenfield (1977) view the doctrine as substituting, in a situation of costly contracting, for a complete contingent contract in which it is efficient to call off the sale in the state of the world where the mistake is discovered. Ayres and Rasmusen (1993), on the other hand, argue that the mutual mistake doctrine, in contrast to unilateral mistake, undermines incentives for information acquisition, and conclude that it is dominated by a suitably limited unilateral mistake doctrine. Indeed, some legal commentators (*e.g.*, Chirelstein, 2006) suggest that in practice the mutual mistake doctrine is used when the court suspects fraud or unilateral mistake but cannot clearly make out the elements of those doctrines on the facts. The doctrine, accordingly, may afford a good opportunity for further theoretical investigation.

Duress and restrictions on contractual modifications. The doctrine of duress denies enforcement to contracts formed under conditions of unacceptable coercion; the standard hypothetical is a contract signed at gunpoint. The key to the doctrine, of course, consists of which conditions are deemed unacceptable. Some courts and commentators have held that difficult economic circumstances can amount to duress; while others suggest that hard bargaining is permitted unless the beneficiary of the promise is responsible for the difficult situation (for example by having isolated the promisor from other potential contract partners) or owes some duty with regard to the difficult situation (for example, by having been granted a local monopoly). Whether a contract is void for duress, however, depends not just on external conditions but also on the content of the contract. For instance, a contract to salvage the cargo of a sinking ship in exchange for 10% of the value of the cargo might be enforced, while one that set a fee of 90% of cargo value probably would not.

A similar functional problem arises in contracts that modify duties created by previous contracts, because many modifications are entered into under circumstances where one of the parties is dissatisfied with the original contract and is threatening to breach it. For example, in the case of *Alaska Packers v. Domenico*, the court held unenforceable a modification that provided a 66% increase in wages to a crew of sailors who threatened to strike on short notice in remote waters.

Much of the legal commentary on this doctrine justifies it on the norm of party autonomy, on the theory that a choice made under coercive circumstances is no choice at all. This justification is difficult to square with the economic viewpoint, however, unless one interprets it as claiming that the harsh circumstances somehow interfere with the exercise of rationality. But in most cases in which the doctrine is applied, rationality does not seem at issue—even in the gunpoint hypothetical, the victim acts rationally in choosing life over money. For this reason, some have argued (*e.g.*, Bar-Gill and Ben-Shahar, 2004) that the doctrine be made unavailable in circumstances where the coercive threat is credible—that is, where the party making it would actually find it in its interest to carry it out if the coerced party does not agree to the exchange in question.

A better economic justification of the doctrine is found in the phenomenon of rent seeking (*e.g.*, Tullock, 2005)—that we wish to discourage investments in coercion or against coercion. In the gunman hypothetical, the rent seeking is obvious, but as Cooter (1982) and Cooter et al. (1982) suggest, a similar argument can apply to ordinary hard bargaining in situations of bilateral monopoly. Even when an exchange is efficient, in the absence of a well-defined mechanism for dividing the gains from trade, the parties may destroy part of the surplus in attempting to influence its distribution.[43] Additionally, in dynamic settings, the prospect of earning such rents may lead to overinvestment or excess entry (cf. Mankiw and Whinston, 1986). Thus, by reducing the set of possible threats, the law can, in principle, limit the available rents and encourage a more efficient equilibrium.

The rent-seeking explanation is especially apposite in the case of contractual modifications, because parties who make specific investments in an initial contract become vulnerable to holdup. In many circumstances, parties will be reluctant to make specific investments in settings in which contract renegotiation is possible; accordingly, it is, in principle, beneficial for them to commit not to renegotiate if they can credibly do so (Aghion et al., 1994).[44] The law of contracts has been unwilling to allow parties to rule out modification, however, perhaps because of concerns for imperfect information and bounded rationality (Jolls, 1997). It has, however, used various strategies over the years to limit modifications that appear motivated by opportunism. In particularly egregious cases, as Shavell (2006b) points out, courts will void the attempted modification on grounds of duress or unconscionability. In less extreme situations, more flexible tools are available. Traditionally, such regulation operated under the formal rubric of the consideration doctrine, on the theory that a modification that unilaterally altered the contract in one party's favor was not a true exchange. The modern approach, however (*e.g.*, Hillman, 1979) requires the modification to pass a substantive test of fairness or good faith. For instance, a modification made in proportionate response to new circumstances, unanticipated at the time of contracting, would generally be enforced, while an outright attempt to rewrite the original terms of the bargain would not. Because it is difficult to identify objective criteria for which circumstances are unanticipated and what sort of adjustment would be proportionate, however, the standard of enforcement is uncertain, so that some incentives for rent seeking do remain.

Undue influence. A fourth standard limitation on contractual freedom is undue influence. This defense applies when one party is deemed to have such influence over the other that the latter lacks the requisite autonomy for contracting. Classic examples include contracts between lawyer and client, physician and patient, caregiver and invalid, and, traditionally, husband and wife.

The undue influence doctrine overlaps in function with the other substantive doctrines listed above. It shares with incapacity the element of bounded rationality, in that the influential relationship may be thought to interfere with the vulnerable party's ability to exercise independent judgment (and as with incapacity, undue influence does not entirely bar the enforcement of a contract, but rather subjects its terms to *ex post* regulation on grounds of substantive rationality and distributional fairness). It also shares with

[43] Although when the bargaining is conducted under symmetric information it is difficult from a game-theoretic perspective to see why surplus would be destroyed (see Hermalin and Katz, 1993, and the discussion in §2.2.2).

[44] Shavell (in press) notes that there is no empirical evidence that parties are able to do so.

fraud the element of asymmetric information (in that the relationship entails a reposition of trust), and with duress the element of rent seeking (in that assent may be given in response to an implicit threat to withdraw the relationship and its associated benefits).

Undue influence is less commonly invoked in the context of person-to-person contracts than the other doctrines discussed above, but it operates as a cornerstone of the law of fiduciary relationships, and thus of much law governing professional relationships and contracts between organizations and their governing agents. For example, the corporate law duty of loyalty, which subjects dealings between corporations and their officers or directors to special scrutiny, is usefully understood as an application of this basic contractual doctrine.

Unconsionability and public policy. Finally, the unconscionability doctrine serves as a sort of all-purpose limitation on contractual freedom applied *ex post* in cases where the court deems the parties' exchange to be sufficiently unfair in either substantive or procedural terms. Some commentators (*e.g.*, Epstein, 1975) have suggested that the doctrine can be justified in economic terms to the extent that it operates as a discretionary extension of the other major affirmative defenses in marginal cases. For instance, the doctrine has been used to protect consumers or unsophisticated small business people from contractual terms that imposed large *ex post* risks and that were buried in fine print or otherwise insufficiently disclosed, which is in some ways analogous to fraud or incapacity. (As noted in the earlier discussion of fraud and non-disclosure, it is not always easy to determine what kinds of statements or omissions should be treated as fraudulent.) The doctrine may also be partially justifiable on the economic grounds of externality (*e.g.*, E. Posner, 1995), to the extent that it prevents contractual creditors from driving their debtors into insolvency, and thus imposing financial obligations on the debtors' families or on the public. Similarly, when viewed in connection with a variety of legal rules that protect debtors in extreme situations, such as bankruptcy law and other statutory limitations on debt collection, the doctrine can be understood as responding to failures in the markets for credit and insurance.

These efficiency rationales cannot entirely account for the actual operation of the doctrine, which sometimes appears merely to reflect the courts' reluctance to hold a hapless party to a bad bargain, or a paternalist concern that a party has undertaken an excessively high risk. The doctrine's implications for freedom of contract, however, should not be overstated, despite its prominent treatment in academic discussions and introductory student texts. Its effect is limited by the requirement that it can only be applied by a judge as opposed to a jury, and by most judges' appreciation that, at least in commercial cases, parties have good reason to take risks and adequate opportunity to obtain insurance. As a practical restriction on the parties' ability to get their bargains enforced in court, accordingly, it is probably rather less important than either the costs of litigation or than the more general psychological tendency of both judges and juries to resolve ambiguous facts in favor of the party with whom they feel the stronger sympathy.

Similarly, courts sometimes decline to enforce contracts on the grounds of public policy, although this concept is less a specific doctrine than a label for a general practice of limiting contractual freedom on a variety of rationales. The main difference between unconscionability and public policy is that the former is based on unfairness, while the latter is based on the idea that enforcement runs counter to the general goals of the legal or judicial system, even though no specific statutory or common law rule contains an explicit ban on the contract in question. For example, contracts to engage in gambling or in ostensibly immoral behavior were traditionally denied enforcement on such grounds, as were contracts in restraint of trade and contracts thought to promote litigation. A prominent modern case applying the concept is the Baby M case, in which the New Jersey Supreme Court refused to recognize a contract which purported to transfer parental rights in exchange for financial consideration. Because of the breadth of this category, probably the best way to summarize it is simply to say that while the principle of freedom of contract is generally respected by the legal system, it remains subject to a vague and implicit set of limits that operate at its margins, on analogy to the limits imposed by other fields such as tort and criminal law. A more extensive discussion of the concept can be found in Farnsworth (2004).

3. Formation of contracts

In many standardized or spot market transactions, parties enter into contracts without any significant amount of bargaining. In more complicated situations, however, a contract typically follows an extended set of communications that can include offers, counter-offers, and other exchanges of information. Such communications are governed by an elaborate framework of legal rules that determine when, how, and on what terms contractual obligations are created.

Compared to other areas of contract law such as the rules governing remedies for breach, there has been relatively little discussion of this doctrinal subject from an economic point of view. In part, this neglect results from the fact that these doctrines are esoteric and little known outside the legal profession. Additionally, legal discussions on the topic tend to focus on the narrow problem of channeling the parties' communications into conventionally recognizable forms, rather than on more direct regulatory concerns.

From an economic viewpoint, however, the law of contract formation is relevant because it influences the outcome of exchange. By attaching consequences to the various acts and omissions that individual bargainers can choose from in a negotiation, legal rules affect the parties' incentives to make and to respond to offers, to exchange information, and to communicate with one another at all.

In this section, accordingly, we survey the various dimensions in which precontractual behavior can affect the efficiency of exchange, and the way in which the law of contract formation affects incentives along those dimensions.

3.1. Pre-contractual behavior

Even before they meet, potential contracting parties make a number of economic decisions that influence the possible gains from future trade. Not all of these decisions, however, will necessarily be made optimally. Some decisions create positive or negative externalities, and others entail relational investments that are vulnerable to holdup. In this subsection, we identify and discuss three overlapping dimensions of pre-contractual behavior: searching for contractual partners, investigating the value of exchange, and making pre-contractual investments.

3.1.1. Searching for contractual partners

In order for parties to be in position to enter into exchange, they must undertake the expense of searching for potential trading partners. Such expenses include advertising, correspondence, travel, and the parties' time. From the perspective of a social planner, one would want the parties to undertake such efforts up to the point where the marginal costs of additional search just outweigh its expected marginal value. While search can be modeled in various ways (see, *e.g.*, Diamond, 1987), under plausible assumptions, the value of an additional unit of search lessens as the quality of the bargain in hand increases. It follows that it is, in general, optimal for a party to search up to some cut-off level of satisfaction, and then stop. This cut-off level will depend on the cost of search, the distribution of information, the expected bargaining game to be played once search is completed, and so on.

The level of search that is privately profitable for an individual buyer or seller, however, is not necessarily the same as the level that would be socially optimal, for two reasons. First, in markets with bilateral search, each person's search efforts provide a positive externality that reduces the search costs of others. Second, to the extent that there are economic rents associated with trading (*i.e.*, if parties buy or sell at prices that diverge from their reservation prices), some amount of search is motivated by the desire to find a better distributional outcome. These effects work in opposite directions, so it is difficult to generalize about what public policies would be optimal in this regard, but it is possible in principle that enlightened regulation could improve social welfare.

For example, §3.3 below explains how various legal doctrines operate to promote the early formation of contractual liability. Such doctrines will thereby affect the level of search, but the direction of the effect is ambiguous. From the *ex post* perspective of parties who have already found each other, such rules should reduce the incentive for additional search, because a party who holds a binding commitment has less need to search; and a party who is bound faces reduced value from search because even if she finds a better bargain, she will still be liable for the first one. But from an *ex ante* perspective, parties may be more willing to undertake the cost of search if they are assured that any trading partners they find will stick with their deal. Which effect is more important depends on the details of the situation.

3.1.2. Acquiring and disclosing information about the value of exchange

In order to conduct exchange, the parties not only must find each other, but they must also determine whether trade is worthwhile. Both the acquisition and disclosure of such information can be costly, and the parties' willingness to incur these costs will depend on whether they can be recouped in subsequent contract bargaining and performance.

As with informational investment more generally, imposing a duty to share information tends to undercut the incentive to acquire it, so there is a potential tradeoff between efficient production of information *ex ante* and efficient use of information *ex post*. In this regard, Kronman (1978a) argued that requiring commodity traders to disclose private information about market value would undercut their incentives to do market analyses and make forecasts; and Matthews and Postlethwaite (1985) showed how requiring manufacturers to disclose product quality could discourage product testing.

Whether information should be produced, of course, depends upon whether it is socially valuable, as opposed to simply having private redistributive value. For that reason, if information acquisition is, on balance, socially wasteful, it is best to impose a disclosure requirement in order to deter rent-seeking (or a tax on acquisition efforts or windfall profits if that is administratively cheaper).

Whether information is acquired or disclosed before exchange, however, will depend not just on regulatory requirements, but also on property rights, the relational nature of the information, and market structure. For example, in Shavell's (1994) model of information acquisition, a party with exclusive rights over a particular item of property has efficient incentives with regard to acquiring information that bears on the common value of that property, because he internalizes that common value when he sells or uses the property. One could imagine circumstances, however, in which a party without formal property rights had similarly good incentives because he held a monopoly position in bargaining and because the information increased his private value for the property, but not the owner's.

Disclosure may also be compelled by the structure of expectations in bargaining. For example, Grossman (1981) presented an influential model of product warranties in which buyers rationally assume the worst about any seller who does not disclose its private information. Sellers with relatively good information are thus led to disclose it in order to avoid the effect of buyer skepticism, and the result is a separating equilibrium in which the private information is revealed.[45] Note that disclosure induced by such game-theoretic incentives can discourage information acquisition every bit as much as disclosure required by legal regulation.

[45] To be precise, information unravels to the point where the cost of credible disclosure equals the difference between the price charged by the pooled low-quality sellers and the price that the highest-quality member of this pool could obtain through disclosure. Observe the unraveling result depends on the disclosure being effective to convey information. See, *e.g.*, Fishman and Hagerty (2003), who show that if the fraction of customers who can understand a disclosure is too low, then mandatory disclosure benefits informed customers and harms the seller.

Even in cases where mandatory disclosure would not adversely affect the incentive to acquire information, it is still necessary to compare the benefits of disclosure against the costs. In many contexts, especially those involving ordinary consumers, the costs of disclosure will include communication costs—for example, the time that it takes buyers to read and process the information; or the potential cost of being distracted from other, more important information. These communication costs have rarely been incorporated into formal economic models, though Bebchuk and Shavell (1991) is an exception. Craswell (2006) provides an informal discussion of these costs.

Finally, it must be noted that in some instances *partial* disclosure of information is worse than no disclosure (Hermalin and Katz, 2006). This is true when partial disclosure facilitates inefficient discrimination. For instance, disclosure of information correlated with on-the-job performance (*e.g.*, health status) could depress the wages of those disclosing the correlate in equilibrium, causing some talented individuals to exit the labor market.

3.1.3. Pre-contractual investment

The previous discussion emphasized that investigation of exchange value can operate as a sunk investment that is vulnerable to rent-seeking in later bargaining. The same is true of pre-contractual investments generally.

Much of the law of contracts is designed to protect investment incentives, and §4 and §5 below discuss the effect of interpretation rules and contract remedies on those incentives. But these rules come into play only after a contract has been formed. Some investments, however, must be undertaken before it is practical for contractual liability to attach. For example, search efforts inherently must precede contractual negotiation (else who would one negotiate with?), and negotiation must precede the formation of a contract. But both search and negotiation are costly, and their costs are at least in part relation-specific. Similarly, other investments can be delayed until negotiations are completed, but their value is reduced in doing so (Katz, 1996c). Suppliers of goods, for instance, can typically reduce their production costs by buying materials when prices are low or by doing advance work when business is slow; and buyers can increase their value from exchange by investing in complementary inputs. But if they wait until they are finished bargaining to begin preparations, many such opportunities will be lost.

It is not desirable to provide complete protection for pre-contractual investments because such protection would lead to excessive reliance. As the parties negotiate, they may discover information that reveals that the intended exchange should not be pursued. If this happens, any relation-specific investments will be wasted. Optimally, the parties should take the risk of wasted investments into account before making them. The rules governing contract formation, accordingly, should ideally be designed to promote optimal reliance at the optimal time, balancing the benefits of productive investment against the costs of waste.[46]

[46] The mechanism for paying compensation for a taking (governmental exercise of its right of eminent domain) must strike a similar balance (Hermalin, 1995).

3.1.4. Strategic behavior in bargaining

In §3.1.2 above, we discussed incentives to produce and share information regarding the value of exchange. But information exchange is not the only dimension of efficiency in bargaining. Negotiating parties can dissipate resources through various types of rent-seeking: by excessive communications, by haggling and stalling, and by hard bidding that risks losing the bargain entirely. One way to deter such behavior is through substantive legal doctrines that limit the kinds of bargains that can be enforced, and thus lessen the temptation for overreaching; for example, Cooter (1982) justifies the doctrine of duress on such grounds. But another way is to create contract formation doctrines that impose the cost of lost bargains on parties who cause them through excessive rent-seeking.

In order to address the problem of excessively hard bargaining, however, it is necessary to develop a clearer understanding of incentives to engage in it. As we discussed in §2.2.2 above, in the *absence* of asymmetric information the parties should be able to reach an efficient bargaining outcome without outside intervention.

Symmetric information is likely the exception rather than rule, however. Complete symmetry assumes symmetry of knowledge about preferences, among other parameters, which is typically not true (see *e.g.*, Farrell, 1987b). If parties bargain under asymmetric information it is well-established (*e.g.*, Samuelson, 1985; Farrell, 1987b; and Schweizer, 1988) that they will often fail to reach a first-best outcome. The theory of mechanism design implies that for any asymmetric information bargaining game, there is, in principle, some mechanism that maximizes the parties' expected bargaining surplus subject to their participation and truth-telling constraints; that is, achieves the second best.[47] But any particular bargaining game may not be the optimal mechanism, so the ultimate efficiency of the bargain could depend on limitations on bargaining imposed by the law on contract formation.

3.2. Avoiding miscommunication

Finally, before we move to a discussion of legal doctrine, we consider the important role played by formation rules in avoiding miscommunication. Negotiating parties need to know when bargaining has been completed and whether a binding obligation has been formed, since if they have inconsistent understandings in this regard, significant value can be wasted. If a party believes that he has entered into a binding obligation when in fact he has not, he may waste resources attempting to perform or may turn down other profitable opportunities. Conversely, if a party is unaware that a contract has been formed, she may fail to make adequate preparations or even incur multiple liability, resulting in breach of contract and associated losses. One of the most important functions

[47] See Laffont and Martimont (2002), Bolton and Dewatripont (2005), or Caillaud and Hermalin (2000b) for introductions to mechanism design.

of contract formation law, accordingly—and the function to which legal scholars have probably devoted the most attention—is to promote clear channels of communication so that the parties may know where they stand.

The law of contract formation pursues this task in two distinct ways. First, it establishes authoritative forms and terms of art that the parties can use when negotiating their agreements. Second, it allocates the risk of miscommunication so as to encourage parties to take optimal precautions to prevent and insure against misunderstandings, in the same way that the law of torts allocates liability for accidents to the least-cost avoider and least-cost insurer. For example, the doctrines of unilateral and mutual mistake, discussed in §2.5.2 above, illustrate how loss allocation rules can strengthen the parties' incentives to avoid misunderstandings and to allocate any remaining risk on their own.

This issue of clarity in communication is most usefully understood in terms of the problem of interpretation, which we discuss at length in §4 below. Determining whether a contract has been formed raises most of the same issues as determining the terms on which it has been formed. For example, one must identify the meaning of phrases that are used in negotiation, reconcile interpretive differences among the parties or between the parties and the court, establish default rules that apply when the parties have left their negotiations open or spoken ambiguously, and so on. We defer consideration of such issues until the next section. Keep in mind, however, the connection between the legal doctrines discussed here and their role in interpretation discussed in §4. For instance, as we discuss in §4.4.2 below, pre-contractual communications often bear on the interpretation of a contract, especially with regard to issues where its text is silent or unclear.

3.3. Legal doctrines addressing contract formation

The body of legal doctrine relating to contract formation is particularly elaborate, and we do not attempt to survey it here. Instead we discuss a number of key doctrines that help address the central economic aspects of pre-contractual behavior: disclosure, investment, and efficiency in negotiation.

3.3.1. Offer and acceptance

The classic distinction between contract law and other sources of legal obligation is that contract is grounded in voluntary agreement. The fundamental principle of contract formation, it follows, is that there must be mutual assent in order to establish a binding contractual obligation. Assent refers not, however, to the parties' private mental states, which clearly cannot form the basis of a public system of enforcement, but to third-party inferences regarding those states as drawn from the parties' actions and statements.

To simplify this problem of inference, lawyers often say that to form an agreement there must be an offer (which evidences one party's assent) and also an acceptance (which evidences the counterparty's assent); and thus this body of doctrine is often referred to as the law of offer and acceptance. It should be recognized that this way of describing the matter is a formalism that allows interpreters to focus on certain stereotypical features of negotiating a contract and abstract from others. In actual contexts it is often difficult or artificial to identify a particular communication that counts as an offer or an acceptance, and some doctrinal systems (such as the law of sales under UCC Article 2) dispense with the effort. In general, however, the traditional common law has followed a highly formalistic approach in this regard, with the result that the treatises and casebooks are filled with a variety of arcane and colorfully named mechanical rules. Such rules include, among others, the "mailbox rule" (*i.e.*, when parties communicate a distance, a contract is formed at the moment an acceptance is dispatched, not when it is received), the "mirror-image rule" (in order to constitute an acceptance, a responding communication must mirror the offer in all relevant respects, else it is deemed a rejection and counter-offer), and the "last-shot rule" (if parties exchange a series of non-matching offers followed by one of the parties commencing performance, a contract is deemed to be formed with the beginning of the performance constituting the acceptance, and with the terms of the contract supplied by the final offer outstanding prior to acceptance). Some of the formalistic rules have received substantial economic attention (see, *e.g.*, Baird and Weisberg, 1982, and Ben-Shahar, 2005, on the last-shot and mirror-image rules), but many others have not.

Over the last half century, such formalistic rules have come under increasing criticism, and more recent developments tend to de-emphasize formality, making it easier to establish liability without complying with prevailing formal conventions. This movement away from formality has had its most important effect on communicative efficiency; and is discussed at greater length in §4.4.1 below. Similarly, and simultaneously, the trend has been for the courts to establish that liability is incurred earlier in the bargaining process (with the mailbox rule providing an early precursor in this regard). This trend is probably most important with regard to incentives for pre-contractual investment, and we take it up in the subsection immediately following this one.

In addition, however, the particular rules of offer and acceptance also influence the transactions costs of negotiation in potentially significant ways. As an illustration, consider the doctrine of silent acceptance (Katz, 1990a, 1993), under which an offeree must respond affirmatively in order to create a binding obligation; silence operates as an acceptance only in special circumstances. The rule can serve to conserve on message costs. Under it, two messages are required to form a contract, while only one is needed in negotiations that do not result in exchange. A converse rule under which an offeree must respond in order to avoid being bound, would require one communication for a contract and two for a rejection. In the usual context where the majority of offers are rejected and rejections and acceptances cost the same to transmit, the common-law rule is efficient. But in settings where acceptance is likely or where acceptances are costlier to transmit than rejections (for instance, because of a mirror image requirement), the rule is inefficient.

Additionally, the common law rule reduces the costs of rent-seeking through opportunistic offers. Under a silent acceptance regime, offerors will have an incentive to propose inefficient exchanges to offerees with high response costs, in the hopes that the offeree will choose to accept a mildly unprofitable contract in preference to incurring the costs of sending rejections. Such incentives explain why negative option plans of the sort offered by mail-order book clubs and telephone companies typically provide buyers with substantial up-front benefits as an inducement to enter the plan.

Finally, based on an extensive survey of the case law, Craswell (1996b) argues that offer and acceptance doctrines are widely used by courts to promote efficient relational investment. In his view, courts are often in position to evaluate *ex post* whether reliance has been efficient, and in such cases, they can provide good incentives by finding a contract when a party relied reasonably in response to a pre-contractual communication, but finding no contract when the party relied either excessively or not at all. Whether Craswell's argument provides a normative justification for courts practices in this regard, however, as opposed to an accurate positive account of what they do in fact, is open to dispute. In the first place, his survey of cases suggest that the fact of reliance is decisive to a finding of liability only in relatively close cases, not in all cases, so the incentive effects of this practice may be limited. Additionally, in order for courts to employ this tool, they must be able to judge the efficiency of reliance in hindsight. The practical limitations of the legal process, the fact that reliance decisions frequently entail nonverifiable actions, and the cognitive biases associated with hindsight (Rachlinski, 1998) may call this assumption into question. Instead, courts may do better to follow the simpler approach of identifying and holding liable the party who is the least-cost avoider for wasted reliance. A model based on this possibility is presented in the following subsection.

3.3.2. Promissory estoppel and analogous doctrines

The general principle of estoppel was developed in the English equity courts as a corrective device to preclude the operation of legal rules that were thought to yield an unjust result in specific cases. The more specific doctrine of promissory estoppel gained influence in the 19th century as a way to soften the formalities of the consideration doctrine, but in the latter half of the 20th century in the United States, it became increasingly used to loosen the formal requirements of offer and acceptance.

The key element necessary to invoke promissory estoppel is relation-specific investment, which lawyers call reliance. Under it, a promisor is precluded from asserting a lack of mutual assent if she made a promise that reasonably can be expected to induce the promisee's reliance, which actually does induce such reliance, and which will result in injustice if not enforced. "Injustice" is admittedly subjective, but in effect it serves to protect reliance that most people would consider reasonable. Excessive or unreasonable reliance is not protected by the doctrine, although there are obvious difficulties in determining reasonableness in this setting. For example, in the case of *Drennan v. Star Paving*, a general contractor used a subcontractor's offer in calculating his bid on a construction job. After the general contractor won the construction contract, the subcontractor attempted to withdraw the original offer on grounds of miscalculation, arguing that under prevailing rules of offer and acceptance, the offer was revocable until the moment of acceptance. The court held that the subcontractor's bid was non-retractable, notwithstanding that there was never any explicit promise to hold it open and that the general contractor could have explicitly bargained for a binding option.

The efficiency of this result, and of the estoppel doctrine generally, depends on whether such liability is necessary to protect investment incentives. Katz (1996c) offers a model of pre-contractual investment in which one party makes an offer and the counterparty chooses to rely; and either of these decisions can be taken over a period of time leading up to the deadline for ultimate performance. Excessively early reliance is inefficient in this model because there is too high a chance that the investment will be wasted; excessively late reliance is inefficient because the productive surplus from investment cannot be enjoyed. The main conclusion of the model is that estoppel liability is desirable if the relying party cannot protect the value of its reliance investment in post-reliance bargaining, but undesirable if it can. The underlying intuition is that if the offering party holds the *ex post* bargaining power, the relying party will refuse to invest unless protected by liability; while if the relying party holds the *ex post* bargaining power, the offering party will refuse to specify its needs for fear of being held liable in circumstances where it does not pay to complete the contract, and being vulnerable to rent-seeking through a last-minute counter-offer in circumstances where it does pay to complete the contract.

Katz's model focuses on the timing of reliance investment, but its underlying intuition also applies to situations in which the extent and type of reliance is in question. As a general matter, reliance investments might be protected through liability rules or by *ex post* bargaining; and in cases where bargaining power is sufficient to provide incentives, legal liability is superfluous and may even provide excessive protection. How these considerations play out in individual bargaining contexts, however, depends on the precise nature and sequence of play. For instance, Bebchuk and Ben-Shahar (2001) consider the possibility of legal rules under which courts can condition liability on the level of reliance investment, on its reasonableness, or on the reasonableness of positions taken in *ex post* bargaining, and show that such rules can be used to promote efficient bilateral incentives. They also discuss the effects of liability rules on the parties' incentives to enter into negotiations initially, and show that the prospect of liability need not deter initial negotiation (although under some conditions, it might). The informational requirements of such rules are of course significant, which limits their practical applicability in many situations.

The idea of conditioning liability on the reasonableness of bargaining behavior, however, underlies a related legal doctrine: the duty to bargain in good faith. This doctrine is much less widely used than promissory estoppel; it is applicable when the parties enter into an agreement that is too indefinite for a court to enforce as written, but that could be enforced if the parties were to undertake an additional round of bargaining in which they filled in enough gaps. For example, in *Teachers Insurance & Annuity Co. v. Tribune Co.*, 670 F. Supp. 491 (SDNY, 1987), the defendant refused to complete a complicated real estate deal, following significant negotiations resulting in a commitment letter that referred to a "binding agreement," after learning that interest rates had significantly dropped and that the deal would probably not qualify for favorable tax treatment. The court held that even though the defendant had reserved a further right of approval on the part of its board of directors, it was obliged to continue negotiations in good faith, which in the specific context meant that it could not condition its approval on a contingency (*i.e.*, favorable tax treatment) that was deliberately not included in the commitment letter, and could not withdraw from the deal simply because interest rate changes had rendered it unprofitable.

The precise contours of the duty to bargain in good faith are not entirely clear. For instance, it is difficult to predict whether a court would hold it bad faith for a party to insist that all open terms be resolved in its favor or to ask that some previously settled term be re-opened, on the grounds that the deal had turned sufficiently sour that a fair distribution of rents justified such an adjustment. (Cases applying the modern doctrine of modification, which also turns on a good faith standard, have often held such adjustments to be acceptable.) The good faith doctrine has been criticized for its uncertainty and apparent subjectivity in application. Nonetheless, the bulk of commentators continue to approve of the doctrine as a means of protecting relational investments in negotiation. In *TIAA v. Tribune*, for instance, the parties had already incurred significant costs and, more importantly, a deal of such complexity could not have been concluded without sinking such costs. Absent some form of pre-contractual liability, it would be difficult to enter into complicated contracts without subjecting reliance investments to the risk of holdup.

Other legal systems have developed analogous rules to deal with this functional problem. For example, the civil law regimes of continental Europe do not use the doctrine of estoppel, but they do include an invigorated version of the duty to bargain in good faith, known as *culpa in contrahendo*. Under this doctrine, parties entering into contract negotiations are held to a mutual duty of care that is intended to protect the reasonable expectations with which they enter bargaining. This duty has traditionally been regarded by comparative lawyers as being more demanding than the implicit duties imposed by the doctrine of promissory estoppel, but one can argue that modern development in the common law have effectively brought the two systems closer together in this regard.

3.3.3. Duty to disclose

Finally, the common law also imposes an ill-defined duty to disclose especially important information when negotiating with contractual partners. The duty does not extend to all information a counterparty might find relevant; rather, it is limited to situations in which nondisclosure would be regarded as effectively equivalent to a representation that the information does not exist. For example, in the classic case of *Laidlaw v. Organ*, a trader with advance knowledge of a peace treaty that would shortly result in the lifting of a naval blockade bought up tobacco that skyrocketed in value when the news became public. When the seller sought to avoid the contract on the grounds of fraud, the court suggested that, under the circumstances, the parties were entitled to hold each other at arms' length and to retain the benefits of such information for their private use. On the other hand, a homeowner who sells a termite-infested house to a buyer who is known to be unaware of any problem may well be under a duty to disclose it, especially if the buyer would not be able to discover the infestation though the exercise of ordinary diligence.

The duty to disclose has received substantial attention in the scholarly literature; and it was recognized early on that nondisclosure might have effects on the fairness and efficiency of exchange. In the law and economics literature, it has been recognized, at least since Kronman (1978b), that imposing a duty to disclose could deter information acquisition. Though later scholarly efforts (see, *e.g.*, the discussion in §3.1.2 above) have refined economic understanding of the issue, doctrinal application of these efforts has been limited. Kronman argued that, at minimum, there should be a duty to disclose information that was acquired without effort or expenditure, on the grounds that such a duty would not reduce the availability of information and would improve the *ex post* efficiency of exchange, but even this suggestion has not made its way into the case law, which has tended in this area to focus more on the question of rights than on efficiency.[48] Still, the trend in legal decisions seems to run in the direction of increased disclosure requirements, especially in settings where the interests of consumers, workers, or small businesses are involved.

3.3.4. Overall assessment of the law of contract formation

The division between the law of contract formation and other parts of contract law is not clear-cut. For instance, some of the rules discussed above in §2.5 on freedom of contract could alternatively be interpreted and analyzed as formation rules. Take for example the Statute of Frauds (see §2.5.1). From a freedom-of-contract perspective, the Statute helps to deter boundedly rational parties from entering into contracts without adequate deliberation, and may help internalize costs that are imposed on the public legal system when parties enter into agreements without adequate evidentiary backing. But it also serves some of the purposes discussed above, such as the disclosure of information or the reduction of miscommunication. Similarly, the mistake doctrine allows inefficient contracts to be avoided, but it also encourages parties to share information that would prevent misunderstandings and the wasteful investments that can result from them.

Additionally, as elaborated in §5 below, the rules governing contract formation significantly interact with the rules governing the remedies for breach. For instance, as we will see, promissory estoppel may substitute for a formal offer and acceptance in the appropriate case, but the damages theoretically available to the disappointed promisee may be rather less than the damages that would be awarded if the proper formalities have been observed (although there is scholarly controversy regarding whether courts

[48] While *complete* disclosure will generally enhance efficiency, *partial* disclosure can reduce welfare (Hermalin and Katz, 2006).

actually follow this theoretical distinction in practice). Again, these doctrinal complications are likely to provide fruitful opportunity for additional economic research in future years, as is the law of contract formation generally.

4. Interpretation of contracts: contractual incompleteness

Offer and acceptance only begins the process of contractual exchange. In order for the transaction to be completed, the contract must be performed, and if performance is not forthcoming, enforced. Probably the most common source of contractual disputes is differences in interpretation, if only because the parties have limited incentive to pursue a dispute if they can foresee and agree upon its likely outcome. The problem of contract interpretation thus provides a central backdrop for the law of contracts, which contains many rules and principles that are designed to address it.

The legal issue of interpretation corresponds to the economic issue of contractual incompleteness—a topic that has been a central focus of research in microeconomic theory in recent years (*e.g.*, Grossman and Hart, 1986; Hart, 1987; Hart and Moore, 1999; and Tirole, 1999). Contractual incompleteness captures the idea that real-life contracting can fail to produce contracts that are as precise and detailed as traditional—albeit possibly naïve—economic theory predicts. Economists typically attribute this failure to an informational asymmetry: the parties to the contract anticipate observing events that they might wish to be contingencies, but which cannot serve as contingencies because they are not verifiable (*i.e.*, observable by a third-party adjudicator of any contract dispute); in other words, the parties will be better informed about payoff-relevant information than any third party. For their part, lawyers often attribute the failure to complete contracts to the inevitable ambiguities in ordinary language or to some bounded rationality on the part of the parties (perhaps arising from their decision not to employ lawyers in the drafting of the contract).

In this section, accordingly, we discuss the problem of contractual incompleteness and relate it to the question of interpretation and to associated legal doctrines. We begin in the next subsection by considering various definitions of contractual incompleteness. Subsection 4.2 then discusses the sources of contractual incompleteness, including *ex ante* determinants present at the outset of contracting, such as bounded rationality and asymmetric information, as well as *ex post* factors such as verification costs and dynamic incentives arising from the prospect of renegotiation. Subsection 4.3 addresses the consequences of contractual incompleteness for the efficiency of exchange; and subsection 4.4 outlines and analyzes the main legal doctrines that govern in this area.

4.1. Modeling incomplete contracts

From a theoretical perspective, it is useful to model a contract as a mapping from *verifiable events* to outcomes. For instance, an insurance contract could contain a provision that related damage to one's car (a verifiable event) to a payment to the insured (an outcome).

In this context, "verifiable" means an event is observable not only by the parties to the contract, but also by any third party (*e.g.*, a judge) who might be called upon to adjudicate a dispute. The focus on verifiable events is motivated as follows. Were an outcome contingent on an unverifiable event (*i.e.*, one not observable to the third party), then there would be no way for the third party to judge the extent of breach of contract (if any) or even who breached (if anyone did). Hence, a contractual obligation that is contingent on an unverifiable event cannot be effectively enforced by a third party.

In the incomplete-contracts literature, it is standard to assume that the parties to a contract can observe events that cannot be verified by any judge. In the parlance of the literature, such events are described as observable, but not verifiable. As we will see, the parties could ideally wish to base their contract on such observable, but unverifiable events.

Enforcement would also be difficult if one of the parties to the contract couldn't observe an event on which an outcome was contingent (for example, as in the case of a consumer who cannot tell whether a mechanic has properly repaired an automobile). That party would not know the extent to which the other party was out of compliance with the contract (if he was). Such ignorance would, thus, make a contract impossible to enforce even through private sanctions of the sort discussed in §5.4.2 and §5.4.3 below. For this reason *observability* is considered a minimal informational requirement for an event to define a contractual contingency.

To be more formal, if we take Ω to be the set of verifiable events (with ω a representative element) and $\mathcal{A}$ to be the set of outcomes, then a contract can be seen as a mapping, $C : \Omega_C \to \mathcal{A}$, where $\Omega_C \subseteq \Omega$. Contractual *in*completeness can, then, be seen as situation in which the parties to the contract would or should ideally wish to base their contract on some set other than Ω_C (for example, Ω if $\Omega_C \subset \Omega$; or the set $\mathcal{O}$ of observable events).

Within the economics literature, the terms verifiable and observable are typically employed without any explicit consideration of the costs associated with investigation, measurement, documentation, or monitoring; that is, activities necessary to make information useful contractually. To an extent, one implicit set of assumptions typically employed is that observable events are observable at no cost, similarly for verifiable events, while *un*verifiable events would be verifiable only at a prohibitive cost (possibly infinite). As Hermalin and Katz (1991) point out, events can be "partially" verifiable in the sense that although a third party cannot observe them with the precision of the parties to the contract, a third party can still observe some information about the event that causes him or her to update his or her beliefs about what happened in a way useful for adjudication (see discussion below in §4.1.3 and §4.3.1). In any case, whatever the costs of observation and verification, they are almost always taken to be exogenous to the model. Endogenizing these costs through the design of measurement or monitoring systems remains, for the most part, an important area for future research.

4.1.1. Literal incompleteness and unmapped contingencies

Following Hermalin and Katz (1993), we make a distinction between literal incompleteness, which we consider now, and economic incompleteness, which we address in §4.1.3.[49]

A contract is literally incomplete if an event or contingency can arise that is not anticipated by the contract; hence, the contract is silent with respect to what should happen given this event or contingency. Literal incompleteness corresponds to a situation in which there are elements of Ω not in Ω_C. Ayres and Gertner (1989) refer to such as elements as "gaps." Other scholars have referred to them as *unforeseen contingencies*. Let $\bar{\Omega}_C$ denote the set of gaps or unforeseen contingencies (*i.e.*, $\bar{\Omega}_C = \Omega - \Omega_C$).

The assumption of literally incomplete contracts has played an important role in law & economics (both implicitly, as in Shavell, 1980 and Rogerson, 1984 and explicitly, as in Goetz and Scott, 1981; Ayres and Gertner, 1989; and Hadfield, 1994). Nonetheless, as Hermalin and Katz (1993) observe, it is a potentially problematic assumption, because it is so easy to complete contracts by adding a stereotypical residual ("none-of-the-above") clause to a contract. That is, literal *completeness* can be achieved simply by adding a clause that states, "if an event (contingency) other than those listed above occurs, then the outcome shall be . . ."

In order to explain literally incomplete contracts, consequently, it is not enough to assume that the parties fail to foresee some contingencies. It would seem necessary to assume also that the parties fail to foresee that they could fail to have foreseen some contingencies.

Alternatively, one might assume the gap is rational, insofar as there is some affirmative reason why the parties deliberately left contingencies unmapped. One reason could be that the parties are content to let the courts apply a particular outcome, and this outcome is no worse than what the parties might have provided on their own. Under traditional common law, for instance, courts often responded to contractual gaps by treating the contract as a nullity, on the grounds that there had been no "meeting of the minds." The result would be that the parties would be left where they lay when the contingency arose (although parties who had partially or fully performed might be entitled to a refund or to restitution of any value conferred on the counterparty). Modern-day courts, in contrast, are more likely to react to a gap by trying to fill it with either an objectively reasonable term, or with their best guess as to what the parties would have wished had they negotiated over the contingency in question. (Although if the gaps are too significant or the parties' hypothetical wishes too unclear, the traditional approach of declaring the contract to be at an end is still a real possibility.)

To the extent that courts apply a predictable default rule in such situations (or even an unpredictable one), one could say that there is no such thing as a literally incomplete contract—rather, the implicit residual clause is just that the parties agree to go to

[49] Ayres and Gertner (1992) refer to the first type of incompleteness as "obligational" and second as "insufficiently state contingent."

court—or follow existing legal requirements that they pursue private arbitration or other method of dispute resolution—and abide by any result that is reached there. At a semantic level, such logic is unassailable, if perhaps unappealing to those worried about truly naïve parties who indeed failed to foresee that they might have failed to have foreseen a contingency.

4.1.2. Linguistic under- and overdetermination

In many situations in which a contract does not point to a clear outcome, the problem is not that the parties have said nothing; rather, it is that they have said too little or too much. For example, the parties might provide multiple and inconsistent provisions dealing with the same event. Alternatively, the parties might use terms that admit multiple meanings or that depend on other terms of the contract.

One could treat such cases as equivalent to contractual gaps, arguing that, if the parties have not settled a term definitively, they have not settled it at all, but this interpretation does not appear to comport with the behavior of either legal institutions or actual bargainers. An alternate approach, accordingly, would be to model contracts not as single-valued functions, but as multi-valued correspondences. Formally, instead of representing a contract as a mapping of the form $C : \Omega_C \to \mathcal{A}$, we could represent it as a mapping of the form $C : \Omega_C \to P(\mathcal{A})$ where $P(\mathcal{A})$ denotes the power set of $\mathcal{A}$.[50] Under this interpretation, incompleteness would arise whenever the contract mapped an event to a set with more than one outcome, so that the person applying the contract would have to choose among those outcomes based on extra-contractual factors.

This type of incompleteness has received less attention in the literature, but some recent work has begun to explore its implications. For instance, a recent paper by Hart and Moore (2004) models a contract as a list of outcomes to which the parties are restricted. The determination of which outcome from that list is implemented occurs through *ex post* bargaining. Through this approach, they show that a relatively loose contract (*i.e.*, a relatively long list) preserves flexibility, allowing the parties to make use of new information that arises, but at the expense of distorting *ex ante* investments due to the increased danger of holdup. Similarly, Ben-Shahar (2004) has argued that courts should respond to such incompleteness by granting both parties the option to enforce the agreement on whatever terms are most favorable to the counterparty; such a response would preserve the benefits of whatever bargain the parties had already reached, while allowing them to enjoy further gains from trade through further negotiation.

A criticism of Hart and Moore (2004) is that their model relies heavily on some strong assumptions. First, they rely heavily on the observable but unverifiable distinction. As Hermalin and Katz (1991) and Maskin and Tirole (1999) note, there are reasons to doubt this distinction in many contexts. More critically, they also rely on the ability of parties to carry out what would normally be seen as incredible threats. Hart and Moore justify

[50] A sigma field should $\mathcal{A}$ be non-denumerable.

these assumptions by appealing to the psychological tendency of people to sacrifice resources in order to get revenge against those who have, in their eyes, mistreated them (see Rabin, 1993, for a discussion of such psychological effects in economic contexts). While such tendencies might exist among individuals, one is hesitant to assume that sophisticated parties (*e.g.*, firms) would be prey to such base emotions.

4.1.3. Economic incompleteness and coarse mappings

Much of the *economic* literature on incomplete contracts has focused on a different standard of incompleteness: A contract is incomplete when the set of verifiable events is not the same as the set of observable events.[51] That is, even if the contract is literally complete (*i.e.*, $\bar{\Omega}_C = \emptyset$) it would be judged *economically incomplete* if $\Omega \neq \mathcal{O}$, where $\mathcal{O}$ is the set of observable events. In this case, the parties would benefit from being able to condition their contractual obligations more finely (for instance, by allowing an excuse when performance is commercially impractical), but they cannot do so because the condition cannot be effectively enforced (for instance, because the court cannot tell the difference between commercial impracticability and mere seller recalcitrance).

As noted earlier, it is natural to assume $\mathcal{O}$ is "larger" than Ω. In the literature, there are essentially two ways this assumption is modeled.[52] One is to define Υ as the set of observable, but unverifiable events (with representative element υ) and take

$$\mathcal{O} = \Upsilon \times \Omega.$$

In other words, an observable event is a vector consisting of events that the parties can observe, but not verify, and events that the parties can both observe and verify.[53] Under this formulation, a verifiable event ω reveals not only that ω has occurred, but also that

$$\upsilon \in \mathcal{O}_\omega \equiv \{\upsilon | (\upsilon, \omega) \in \mathcal{O}\}$$

(the set $\mathcal{O}_\omega$ is called the *section* of $\mathcal{O}$ *at* ω). Because, absent additional assumptions, $\mathcal{O}_\omega = \Upsilon$ for all $\omega \in \Omega$, it might at first seem that knowing that υ is in $\mathcal{O}_\omega$ is not useful information. But if some (υ, ω) pairs are impossible, then some sections satisfy $\mathcal{O}_\omega \subset \Upsilon$ and, therefore, learning ω can lead to inferences about which υ occurred.

[51] Obviously, both sets must be defined in some way over *relevant* events; that is, we don't want to say the sets differ if the observable, but unverifiable events are irrelevant to the contracting situation. Defining relevance is, however, difficult insofar as one strain of the literature (*e.g.*, Hermalin and Katz, 1991; Maskin and Tirole, 1999; and Edlin and Hermalin, 2001) has been devoted to showing when the observable, but unverifiable distinction is unimportant; that is, situations in which the parties can achieve the same outcomes with incomplete contracts as they could with complete contracts. In such situations, it would be misleading to say that the observable and verifiable sets are the same over relevant events, because equilibrium play, but not outcomes, would be different were it feasible to base contracts on the set of observable events. A definition of relevant events must, therefore, account for potential differences in the ensuing game, as well as outcomes.

[52] An example of the first way is Grossman and Hart (1986). An example of the second way is Anderlini and Felli (1994).

[53] It is worth noting that both Υ and Ω could, themselves, be vector spaces.

In such situations, a so-called "forcing" contract (as in Harris and Raviv, 1978) could be useful. Suppose the parties wish to induce one party, "the actor," to choose a specific $\hat{\upsilon}$. If there are ω such that $\hat{\upsilon} \notin \mathcal{O}_\omega$, then those ω constitute proof that the actor has not chosen the desired action and the contract can, therefore, threaten the actor with sufficient punishment should such ω occur that he would never choose an $\upsilon \in \mathcal{O}_\omega$. Hence, some undesirable actions can be avoided. Ideally, if there is a subset $\Omega_0 \subset \Omega$ such that, for all $\upsilon \neq \hat{\upsilon}$, $\upsilon \in \mathcal{O}_\omega$ for some $\omega \in \Omega_0$, but $\hat{\upsilon} \notin \mathcal{O}_\omega$ for any $\omega \in \Omega_0$, then $\hat{\upsilon}$ can be achieved by a contract that sufficiently punishes the actor for any $\omega \in \Omega_0$ and rewards him only for $\omega \in \Omega \backslash \Omega_0$.[54]

Even if all (υ, ω) pairs are possible—so $\mathcal{O}_\omega = \Upsilon$—the *conditional distribution* of υ given $\upsilon \in \mathcal{O}_\omega$ may differ from the distribution given $\upsilon \in \mathcal{O}_{\omega'}$, $\omega \neq \omega'$. Such differences in distributions can be very powerful, often allowing contractual solutions that replicate the outcomes that would have occurred were the parties able to contract on $\mathcal{O}$ directly (see discussion in §4.3.1; also Hermalin and Katz, 1991).

The second way economic incompleteness can be understood is to take Ω to be a *partition* of $\mathcal{O}$; that is, each ω is a subset of $\mathcal{O}$, any given element of $\mathcal{O}$ can be in only one ω, and the ωs, as a class, contain all the elements in $\mathcal{O}$.[55] Of course, to make this interesting, there must be at least one ω with two or more elements of $\mathcal{O}$ in it; that is, for this ω at least, learning ω does not perfectly reveal the observable information. Conversely, economic *completeness* would correspond to a situation in which each ω contained a single element of $\mathcal{O}$.

The standard interpretation of this second representation of economic incompleteness is that there is inherently no difference between what is observable and what is verifiable, except that it is impossible or prohibitively expensive to describe the observable events with sufficient precision in a contract.[56] Instead, the observable events can only be described coarsely; so that different observable events get lumped together in a single ω. For example, consider the "contract" between us, the authors of this chapter, and the editors of this volume. Without effectively writing the chapter themselves, it is impossible for the editors to stipulate fully what the chapter should look like; the best they can do is stipulate the topics to be covered and the overall length of the chapter.

[54] The notation $\mathcal{S} \backslash \mathcal{T}$ denotes the set whose elements are in $\mathcal{S}$ but not $\mathcal{T}$ (*i.e.*, $\mathcal{S} \backslash \mathcal{T} = \mathcal{S} \cap \mathcal{T}^{\mathrm{c}}$).

[55] Formally, the ωs satisfy

$$\omega \cap \omega' = \emptyset, \forall \omega \neq \omega'; \text{ and } \bigcup_{\omega \in \Omega} \omega = \mathcal{O}.$$

This second way of representing economic incompleteness can be seen as a special case of the first in which $\mathcal{O} \subset \Upsilon \times \Omega$ such that $\mathcal{O}_\omega \cap \mathcal{O}_{\omega'} = \emptyset$ for any two ω, ω' and $\bigcup_{\omega \in \Omega} \mathcal{O}_\omega = \Upsilon$.

[56] Typically these costs are taken to be "writing costs." They could also, as Rasmusen (2001) notes, be due to "reading" costs.

4.1.4. An illustrative dispute

Each of these models of contractual incompleteness has some appeal, though none is entirely satisfactory as a complete account of the phenomenon. To illustrate, consider the case of *Spaulding v. Morse*, in which a divorcing couple entered into a maintenance agreement under which the husband agreed to pay child support of $1200 per year until the couple's son entered some "college, university, or higher institution of learning . . .," and $2200 per year for four years thereafter. The dispute arose when the son completed high school and was immediately drafted into the US Army, and the husband took the position that his contractual obligation was suspended while the son remained in military service. One can view this dispute as arising from literal incompleteness if we say that the contract simply did not provide for the case in which the son was drafted into the army. In order to take this view, however, we must read the term in question in a non-literal way, since a literal reading makes it appear to be a residual clause. (That is, it appears to say that so long as the son never started college—*e.g.*, if he chose to pursue a military career or even if he were killed in combat—the husband would be required to make yearly payments in perpetuity.)[57]

One can alternatively view the dispute as arising from linguistic underdetermination if we focus on the phrase "child support." Does the phrase simply refer to any payments made for the benefit of the son, or is it instead limited to payments that are objectively necessary for the son's adequate maintenance? Ordinary language is probably not precise enough to provide a definitive answer to this question, and the best response may be to say that the phrase can mean either or both.

Finally, we could view this dispute as a case of economic incompleteness if we assume that what the parties actually wished to do was to condition payment on the son's need for maintenance (an observable but perhaps not verifiable event) but that they instead conditioned on the son's not yet having completed four years of college (a coarser but perhaps more easily verifiable proxy). Had the parties been able to describe the former condition at reasonable cost in their divorce settlement, the whole dispute could have been avoided. On the other hand, it is not clear that the couple had such a clear purpose in arriving at the clause in question. Another possibility is that the husband simply wished to pay as little as possible, the wife wished to receive as much as possible, and the clause was a rough compromise.

As the example indicates, modeling contracts as a formal mapping from events to outcomes is useful, but it could fail to capture two important real-world features of the situation—namely, that parties do not always possess a well-specified understanding of the domain of possible events, either *ex ante* or *ex post*, and that they write contracts in ordinary language with all its ambiguities, rather than in the formal but restricted

[57] Nobody in the case made this argument, however, not even the wife. Instead, the court found in favor of the husband, on the grounds that the unstated purpose of the clause was to provide financial support for the son in his minor years, and once the Army had taken responsibility in this regard, the purpose had been served.

language of mathematics. This is not to say, however, that such failures necessarily invalidate or diminish the insights of formal modeling. Nor is it to say that formal modeling can't deal with these features should they be important. Certainly, the first could be readily modeled by incorporating some uncertainty over what the domain of events might be. The second is potentially trickier, but techniques such as assuming asymmetric information with regard to the meaning of the terms, uncertainty over their interpretation, or even suspension of the usual common-priors assumption could be useful. With all that in mind, the next subsection discusses possible economic causes of contractual incompleteness.

4.2. The sources of contractual incompleteness

4.2.1. Bounded rationality

A common explanation for incomplete contracts, especially literally incomplete contracts, is bounded rationality.

The simplest "model" of bounded rationality is that people make mistakes. They fail to foresee all possible contingencies and, thus, their contracts suffer from unforeseen contingencies. However, as noted above (§4.1.1), failure to foresee certain contingencies need not generate incomplete contracts unless the parties also fail to foresee that they may fail to foresee certain contingencies. If the parties recognize their imperfect foresight, then they can complete any contract with a residual ("none-of-the-above") clause.

Of course, the fact that the parties *could* complete any contract with a none-of-the-above clause does not mean that they would wish to do so. After all, they may fear that the ideal response to different unforeseen contingencies varies with those contingencies. By its nature, a none-of-the-above clause is a one-size-fits-all solution. Hence, the parties may wish for flexibility should an unforeseen contingency occur; legal proceedings can provide such flexibility insofar as the court's ruling will typically depend on the contingencies that have occurred. How the courts should respond to these "intentional" gaps has become an issue in the literature (see, *e.g.*, Ayres and Gertner, 1989, 1992). We discuss this literature in §4.4 below.

To an extent, deciding how the courts should respond depends on the reason for unforeseen contingencies in the first place. As, however, E. Posner (2003a) points out, economic theory does not offer particularly good or widely accepted means of modeling unforeseen contingencies (or bounded rationality more generally).[58] To the extent models of unforeseen contingencies exist, they are too abstract to be readily utilized in the study of contracts. Admittedly, one could simply appeal to the existence of gaps in actual contracts to justify assuming literal incompleteness. Such an atheoretic rationale,

[58] For a general survey of recent modeling on bounded rationality see Rubinstein (1998). For a more specific survey on unforeseen contingencies see Dekel et al. (1998) (also see Dekel et al., 2001).

however, makes it difficult to analyze important issues such as how the courts would know whether gaps are truly mistakes or have arisen because one side or the other is being strategic. This, in turn, affects how the courts should respond to gaps (see Ayres and Gertner, 1989, 1992, and Shavell, 2006a).

Although some parts of the literature appeal to unforeseen contingencies as a possible motivation for *economic* incompleteness (see, *e.g.*, Maskin and Tirole, 1999), it is difficult to reconcile such bounded rationality with the "dynamic-programming" rationality (to use Maskin and Tirole's term) that the incomplete-contracts literature typically assumes. Other, more rational, explanations for economic incompleteness strike us as more plausible. We explore some of these explanations in the next few sections.

4.2.2. Describability and contracting costs

As we observed, a complete contract between us and the book's editors about what this chapter should be is rather impractical. If the editors spelled out in great detail what they wanted, they would, in effect, be writing the chapter themselves. Their costs of doing so presumably outweigh whatever benefits such detailed contracting would generate. More generally, there is a cost to writing fine-tuned contracts. In some instances, such as our chapter example, it is simply impractical to describe the relevant contingencies finely enough.

Describability. Because describing something can be viewed as an algorithm, it follows from the mathematics of computability (see, *e.g.*, Boolos et al., 2002) that it is conceivable that some events or contingencies cannot be described, in the sense that no algorithm for describing them would ever finish. This point is made by Anderlini and Felli (1994). An intuitive sense of this idea can be had by recalling the well-known math fact that the constant π cannot be given a decimal representation; no algorithm could ever complete the task of fully describing π as a decimal number.

Indescribability *per se*, however, seems a doubtful explanation for economically incomplete contracts for at least two reasons. First, the parties might be able to approximate their desired description arbitrarily well (*e.g.*, 3.1416 might be a good enough approximation of π); therefore, they suffer arbitrarily little from the indescribability of a contingency. In fact, Anderlini and Felli prove just such an approximation theorem. Second, the parties to the contract need to have internal descriptions of the relevant events or contingencies to know if they have occurred; hence, if they cannot write descriptions, it is hard to understand how they know the relevant descriptions. In other words, if we take the feasible contractual contingencies to be subsets of some partition of the relevant state space, then why isn't the set of observable contingencies this same partition?[59] And, if it is, then there is no practical difference between observable and verifiable; that is, the contract is *not* economically incomplete.

[59] Let Σ be the state space. What can be described, the verifiable events, are subsets of Σ (*i.e.*, $\omega \subset \Sigma$ for all $\omega \in \Omega$, where $\bigcup \omega = \Sigma$ and $\omega \cap \omega' = \emptyset$). But if, as seems reasonable, what the parties can describe in writing

Description costs. A better way to proceed is to assume that the relevant states can be described (or described arbitrarily finely), but to recognize that the costs of doing so can sometimes be so large as to render detailed descriptions impractical (see, *e.g.*, Dye, 1985). The chapter-writing example given earlier is an example of such a situation.

Having recognized that contract costs are increasing in the details of the contract,[60] the next question is how to balance the marginal gain from more details against these marginal costs? As it turns out, however, in many settings the marginal *benefit* of adding details is zero; as Hermalin and Katz (1991) and Maskin and Tirole (1999) show, even rough descriptions can be sufficient for the parties to do as well as they could were it practical to write very detailed contracts. We return to this point later.

Another reason the marginal benefits to the parties could be small is that the courts will fill in or interpret the missing or vague terms (see, *e.g.*, Ayres and Gertner, 1989, 1992; or Shavell, 2006a). Hence a model of how the courts enforce incomplete contracts is essential to an analysis of the effort that parties should make to make their contracts more precise.

Evaluation and expertise. It is possible that even if a judge, or other adjudicator of contractual dispute, can observe a relevant variable, he or she may not properly evaluate it. His or her evaluation could differ from that of the parties to the contract, perhaps because of a lack of expertise in the relevant field. This idea can be captured by assuming that, when the parties observe x, the judge observes (concludes) $x + \varepsilon$, where ε is the judge's error in evaluation. The error, ε, could also be introduced because the judge reads or interprets the contract differently than the parties intended. (The model, which we consider in §4.4.1 *infra*, on form versus substance also touches on this point.)

While such errors might, at first, seem to render the relevant variable essentially unverifiable, the parties can, in some circumstances, do as well as they would were the judge to observe the relevant variable without error (Hermalin and Katz, 1991). We return to this point later in §4.3.1.

4.2.3. Complex environments

Segal (1999a) suggests that economically incomplete contracts can arise when the contracting environment is complicated. In his model, the parties wish to trade one "widget" from a set of different widgets. As the number of potential contingencies (*i.e.*, widgets in the set) rises, the optimal second-best contract does increasingly worse. In the limit, the

is the same as what they can describe to themselves, then each ω is, then, an element of $\mathcal{O}$ and $\mathcal{O} = \Omega$. [In a sense, this is reminiscent of Wittgenstein's (1922) Proposition 7 that what we can't think about, we can't verbalize.]

[60] One can model contracts being more detailed by assuming that there is an order, $\succ$, on partitions of $\mathcal{O}$, such that $\Omega_i \succ \Omega_j$ (*i.e.*, a contract based on Ω_i is more detailed than one based on Ω_j) if for each $\omega_i \in \Omega_i$ there exists an $\omega_j \in \Omega_j$ such that $\omega_i \subseteq \omega_j$ and for at least one $\omega_j \in \Omega_j$ there exist ω_i and ω_i' in Ω_i such that $\omega_i \cup \omega_i' \subseteq \omega_j$.

optimal second-best contract does no better than the simple incomplete contract (equivalent, in this model, to no contract). The reason for this worsening performance is that, as the number of contingencies rises, the number of incentive constraints increases. As in all optimization problems, as the number of constraints increases, the optimality of the solution decreases (at least weakly). Given that writing complex contracts is costly, there is a cutoff point at which the environment is so complex that the simple, less costly contract is superior to a complex contract.

Segal's model, while elegant, rests on a number of strong assumptions. Among these is the assumption that almost no variable of interest is verifiable, despite many of them being observable (his Assumption 2, page 60). If either the optimal widget to trade is known *ex ante* or can be *verified ex post*, the model collapses. As we have observed, the assumption of observable but unverifiable can be a questionable one, especially in many of the contexts to which this model is intended to apply.

4.2.4. *Asymmetric information*

As noted in §2.3.2, informational asymmetry at contracting can lead to distortions in the contract that is written. Spier (1992) builds on this idea by suggesting that one consequent distortion could be contractual incompleteness.

A simple model can serve to illustrate her idea. Suppose a manufacturer needs a part for one of its manufacturing machines. The manufacturer can have the part sent by a default carrier, in which case the part will certainly arrive in two days. Normalize the default carrier's price to be 0. Alternatively, the manufacturer can contract with an express carrier. If the express carrier makes no special efforts, the part will arrive in one day with probability $1 - q \in (0, 1)$; with probability q it arrives the day after that. For simplicity, take the express carrier's cost when it makes no special efforts to be 0.[61] Alternatively, by spending $c > 0$, the express carrier can ensure the part arrives in one day. Let $L > 0$ be the amount per day that the manufacturer loses from the idle machine. If the manufacturer selects the default carrier, its payoff is $-L$. If it selects the express carrier, but no special efforts are made, its expected gross payoff is $-qL$. If it selects the express carrier and special efforts are made, its expected gross payoff is 0; but, in this case, the express carrier's gross payoff is $-c$. Assume that there are two types of manufacturer, 0 and 1, with

$$qL_1 > c > qL_0.$$

Observe that it would be welfare maximizing to have the express carrier make special efforts for a high-value manufacturer (type 1), but not for a low-value manufacturer (type 0). Let $\theta \in (0, 1)$ be the probability that the manufacturer is type 1. Hence,

$$\mathbb{E}L = \theta L_1 + (1 - \theta)L_0.$$

[61] Making it a positive amount does not alter the analysis materially, while setting it to zero simplifies the notation.

Assume that the manufacturer's type is its private information.

To close the model, we need to make some assumptions about bargaining between the manufacturer and the express carrier. Assume that if the express carrier can't infer the manufacturer's type, then the express carrier and manufacturer set the price through Nash (1950) bargaining over the expected surplus, $\mathbb{E}L - q\mathbb{E}L$. In that case, the price is

$$p_u = \frac{1}{2}(1-q)\mathbb{E}L.$$

Assume the parameters are such that

$$-qL_0 - p_u > -L_0$$

(this clearly holds for low values of θ). Because it is welfare reducing for a low-value manufacturer to request special efforts, we can assume that the express carrier interprets any request for special efforts to be from a high-value manufacturer. In this case, total surplus is $L_1 - c$. If bargaining is again Nash bargaining, then the resulting price is

$$p_i = \frac{L_1 + c}{2}.$$

Finally, consider a high-value manufacturer who is deciding between asking for special efforts—and thus revealing its identity—or pooling with low-value manufacturers and not asking. Its expected net surplus in the first case is $-p_i$. Its expected net surplus in the second case is $-qL_1 - p_u$. It is readily shown that there exist parameter values such that latter exceeds the former (*e.g.*, $c = 2$, $L_1 = 5$, $L_0 = 3$, $q = 1/2$, and $\theta = 1/4$ would work). Hence, it is possible that asymmetric information leads to a form of contractual incompleteness: Although it would be welfare improving for a high-value manufacturer and the express carrier to have a contingency in their contract requiring the carrier to make special efforts, they don't have such a contingency in equilibrium.

4.2.5. Verification costs

In much of the contracting literature, it is assumed that, if information is verifiable, it can be verified costlessly. In real life, however, information is made verifiable at a cost (*e.g.*, surveillance monitoring, auditing, and record keeping). If such costs are large relative to the benefits that complete contracting affords, the consequence will be incomplete contracts.

There is a strand of the literature on "costly state verification," starting with Townsend's (1979) seminal article, that considers, in part, the consequences of an inability to verify payoff-relevant variables due to the cost of verification. The principal concern of this literature has been on financial contracting; particularly the decision to use debt rather than equity.[62]

[62] See, for instance, Webb (1992) or Hart and Moore (1998).

4.2.6. *Dynamic inconsistency with respect to renegotiation*

A common, although not universal, view in the economics literature is that parties never "leave money on the table" even on out-of-equilibrium paths. More precisely, should the parties ever reach a point at which it is common knowledge that the allocation dictated by the contract is Pareto inferior, the parties will renegotiate the contract. This ability to renegotiate undermines the value of contracts that rely on Pareto-inferior allocations as threats should one or both parties deviate from their contractual obligations. In the lingo of the literature, such contracts fail to be *renegotiation proof*. That contracts be renegotiation proof is, therefore, often a requirement imposed in the literature.

To illustrate the concept of renegotiation-proofness and how it can lead to incomplete contracts, we briefly review the Fudenberg and Tirole (1990) model of renegotiation in agency. In a standard agency model, to induce the agent to work hard, the principal and agent agree to a contract that makes the agent's compensation contingent on the outcome (*e.g.*, salespeople's compensation is frequently tied to their sales). Such a contract is, however, generally not first-best optimal because it causes a risk-averse agent to bear risk. Fudenberg and Tirole consider what happens in the standard model if there is a period after the agent has committed his effort but before its outcome is known (*e.g.*, after he is back from his sales calls, but before the orders arrive) during which the principal and agent can renegotiate. Observe that if, as is typically assumed, the principal is risk neutral, then it is common knowledge that leaving the agent with risky compensation is inefficient; the parties should renegotiate to a fixed, non-contingent wage for the agent. However, a rational agent would forecast such renegotiation and, thus, that his compensation will ultimately be unrelated to his effort. But if it is unrelated to his effort, then he has no incentives to work hard. While Fudenberg and Tirole show that there are ways to restore *some* incentives, the potential for renegotiation limits their effectiveness. Moreover, a consequence could well be that the parties forgo using an incentive contract even though they would use one were renegotiation not possible.

This last point shows how the threat of renegotiation can lead the parties to use less complete contracts than they would were renegotiation infeasible. That is, while they might wish to write a contract contingent on the outcome, they can't. Schwartz and Watson (2003) consider the relation between contractual complexity and the feasibility of renegotiation in greater depth, with a particular emphasis on law & economics issues.

4.3. *Consequences of contractual incompleteness*

4.3.1. *Does incompleteness matter?*

A debate has emerged in economics as to whether incomplete contracts matter. Specifically, even if it is accepted that contracts are incomplete, are there ways for the parties to effectively contract around the incompleteness?

Hermalin and Katz (1991) offer one model in which incompleteness doesn't matter. Let Υ denote the set of observable, but unverifiable variables. Assume that Υ is finite.

Let Ω denote the set of verifiable variables. Take it to be finite as well, with N elements indexed by n. Let there be two parties to the contract: the agent and the principal.

The agent chooses $\upsilon \in \Upsilon$. The principal has preferences over the υs. Suppose that the agent is willing to sign a contract with the principal if his equilibrium (expected) utility is at least some reservation level, which we can normalize to be zero. Let w denote the agent's gross utility and let $w - d(\upsilon)$ denote his net utility ($d : \Upsilon \to \mathbb{R}_+$). Assume it costs the principal $c(w)$ if the agent receives gross utility w, where $c(\cdot)$ is increasing and strictly convex.[63] Finally, assume that the conditional distribution of the verifiable variable, ω, depends on the υ chosen; specifically, let $\pi_n(\upsilon) = \Pr\{\omega_n|\upsilon\}$ and let

$$\boldsymbol{\pi}(\upsilon) \equiv \bigl(\pi_1(\upsilon), \dots, \pi_N(\upsilon)\bigr)$$

be the density over Ω induced by υ.

Suppose that the principal wants the agent to choose a specific υ^*. If we assume that the principal has all the bargaining power, her ideal would be a contract in which the agent agrees to choose υ^* in exchange for a net utility of 0, which costs the principal $c(d(\upsilon^*))$. Except in the special case where υ^* minimizes $d(\upsilon)$, this ideal contract is infeasible because υ is not verifiable and the agent prefers an υ other than υ^*.

A feasible alternative is a contingent-compensation contract; the principal agrees to provide utility level w_n should verifiable outcome ω_n occur. Let $\mathbf{w} = (w_1, \dots, w_N)$. Such a contingent contract will be agreeable to the agent and induce him to select υ^* if

$$\boldsymbol{\pi}(\upsilon^*) \cdot \mathbf{w} - d(\upsilon^*) \geq 0 \qquad \text{and} \tag{7}$$

$$\boldsymbol{\pi}(\upsilon^*) \cdot \mathbf{w} - d(\upsilon^*) \geq \boldsymbol{\pi}(\upsilon) \cdot \mathbf{w} - d(\upsilon) \quad \text{for all } \upsilon \in \Upsilon\backslash\{\upsilon^*\}. \tag{8}$$

Condition (7), the individual rationality constraint, is the condition that the agent agree to the contract. Condition (8), the set of incentive compatibility constraints, is the condition that the agent prefer choosing υ^* to any other υ.

Suppose the principal has all the bargaining power. It is a well-known result (see, *e.g.*, Grossman and Hart, 1983) that if a contract (a vector $\mathbf{w}$) exists satisfying expressions (7–8), then there exists one such that (7) holds as an equality. Because that contract minimizes the principal's expected cost, she will offer that contract. Under general conditions (see, *e.g.*, Grossman and Hart, 1983), $\operatorname{Var}(w|\upsilon^*) > 0$. Hence, by Jensen's inequality, $\mathbb{E}\{c(w)|\upsilon^*\} > c(d(\upsilon^*))$; that is, even if the principal can devise a contract to implement υ^*, it will cost her more than the ideal contract would were υ verifiable. It would seem, therefore, that there is an adverse consequence to υ not being verifiable.

This last conclusion, however, need not follow if the parties have time between the realizations of υ and ω to renegotiate the contract. Suppose that the principal retains all the bargaining power in renegotiation.[64] Because, here, the principal observes υ, the problem identified by Fudenberg and Tirole (1990) is not triggered by renegotiation (see

[63] A natural interpretation is that the agent is risk averse, while the principal is risk neutral, in which case $c(\cdot)$ is the inverse function of the agent's utility of income function.

[64] Edlin and Hermalin (2001) extend the analysis to a broader set of bargaining games in renegotiation.

discussion in §4.2.6). Because the principal's cost increases with the variability of w, the principal would offer, in renegotiation, a non-contingent w, $\bar{w}$. Because the agent's cost from selecting υ is sunk at the point of renegotiation, he will accept $\bar{w}$ if and only if $\bar{w} \geq \boldsymbol{\pi}(\upsilon) \cdot \mathbf{w}$. As she has all the bargaining power, the principal will choose the lowest such $\bar{w}$, namely the one that just equals $\boldsymbol{\pi}(\upsilon) \cdot \mathbf{w}$. Hence $\bar{w}$ is a function of υ, which we indicate by writing $\bar{w}(\upsilon)$. By Jensen's inequality, $c(\bar{w}(\upsilon)) \leq \mathbb{E}\{c(w)|\upsilon\}$; that is, renegotiation lowers the principal's cost. Moreover, because $\bar{w}(\upsilon) = \boldsymbol{\pi}(\upsilon) \cdot \mathbf{w}$, it is readily seen that anticipation of such renegotiation does *not* change the individual rationality nor incentive compatibility constraints (7–8). Moreover, from (7), we see that $\bar{w}(\upsilon^*) = d(\upsilon^*)$. We therefore have

PROPOSITION 6. *Suppose that renegotiation is possible. If a contract* $\mathbf{w}$ *exists that satisfies individual rationality and incentive compatibility (i.e., expressions 7 and 8), then even though* υ *is unverifiable, the principal can achieve the same outcome as she could if it were verifiable.*

In other words, in this context, the distinction between observable and verifiable is not relevant to the outcome.

Proposition 6 rests on the existence of a $\mathbf{w}$ satisfying expressions (7–8). Fortunately, the conditions for such a $\mathbf{w}$ to exist are rather minimal. Somewhat loosely, it is sufficient that the distribution over ω induced by υ^* be distinct from the distributions induced by the other υ.[65]

Proposition 6 also rests on the supposition that renegotiation is possible; that is, that there be a lag between when υ is chosen and when ω is realized. Renegotiation would seem possible in many contexts of interest. For instance, if ω is a judge's observation and, thus, a ruling as to the value of υ, then the parties presumably have time to renegotiate prior to the judge's verdict.

A provocative way to summarize this is as follows: Even when critical variables are not verifiable, parties can often present some relevant evidence that is informative, in a noisy way, about the critical variables. If renegotiation is feasible, then, by contracting on this relevant evidence, the parties can typically write contracts *as if* the critical variables were verifiable.

Renegotiation *à la* Hermalin and Katz is a particular type of game or mechanism. If a larger set of mechanisms is considered, one would find that the observability-verifiability distinction is irrelevant under an even wider set of circumstances. In a series of articles, Maskin and Tirole do just that, considering mechanisms more generally (see, *e.g.*, Maskin and Tirole, 1999; see, also, Tirole, 1999, for an overview).

[65] Formally, it is sufficient that $\boldsymbol{\pi}(\upsilon^*)$ not be an element of the convex hull of $\{\boldsymbol{\pi}(\upsilon)|\upsilon \in \Upsilon\backslash\{\upsilon^*\}\}$. As Hermalin and Katz (1991) discuss, this condition is readily met in most situations of interest. As Edlin and Hermalin (2001) show, for more general bargaining games in renegotiation, a slightly stronger condition could be required: there exist a subset Ω^* of Ω such that $\Pr\{\omega \in \Omega^*|\upsilon^*\} > \Pr\{\omega \in \Omega^*|\upsilon\}$ for all $\upsilon \in \Upsilon\backslash\{\upsilon^*\}$.

In a mechanism, the parties send messages to the mechanism arbitrator (possibly the judge or other dispute adjudicator) about the variables they can observe. As Farrell (1987b) notes, perhaps the earliest recorded use of a "mechanism" is the one King Solomon used to determine which of two women claiming to be the mother of a child was the true mother.

A review of mechanism design is beyond the scope of this chapter. We can, however, give a sense of its application in this context. Suppose that there is some task that either of two parties, 1 or 2, could do. *After* contracting, the parties observe their costs $(c_1, c_2) = \upsilon$ of doing the task, but cannot verify them. Suppose that $c_i \in \{L, H\}$, where L (low) is smaller than H (high).

Efficiency requires that the party with the lower cost do the task. The parties might further wish to have the one who doesn't do the task partially compensate the other (*i.e.*, pay some portion of the cost). But, by assumption, the parties cannot write such a contract directly because it would be contingent on the realizations of their costs. Fortunately, however, a mechanism can be constructed that replicates the desired outcome: Both parties simultaneously announce their individual costs (*i.e.*, 1 announces c_1 and 2 announces c_2). The following rules govern what happens next:

- If they announce the same cost, then a coin is flipped. The loser of the coin flip has to do the task, but is paid $\hat{c}/2$ by the winner, where $\hat{c}$ is the commonly announced cost.
- If they announce different costs, then the lower-cost party must do the task, but she is paid $H/2$ by the high-cost party.

If the parties announce truthfully, then the allocation of the task will be efficient. It remains, however, to check that the parties will announce truthfully in equilibrium (*i.e.*, that the mechanism is incentive compatible). There are three cases to consider. In each, we are assessing whether truth telling is a best response to truth telling; and, therefore, does a truth-telling equilibrium exist.

1. $\upsilon = (L, L)$. If a party announces truthfully, she expects to get

$$\frac{1}{2}\left(\frac{L}{2} - L\right) + \frac{1}{2}\left(-\frac{L}{2}\right) = -\frac{L}{2}.$$

   If she lies, she expect to get $-H/2$; hence, she does better to tell the truth than to lie.
2. $\upsilon = (H, H)$. If a party announces truthfully, she expects to get $-H/2$ (see calculation above). If she lies, she must do the task, but is paid $H/2$; hence, her net is $-H/2$. She does least as well to tell the truth; that is, truth telling is a best response.
3. $\upsilon = (L, H)$ or (H, L). Consider the higher-cost party. If she tells the truth, she expects to get $-H/2$. If she lies, she expects to get

$$\frac{1}{2}\left(\frac{L}{2} - H\right) + \frac{1}{2}\left(-\frac{L}{2}\right) = -\frac{H}{2};$$

so truth telling is a best response. Consider the lower-cost party. If she tells the truth, she expects to get $H/2 - L$. If she lies, she can expect

$$\frac{1}{2}\left(\frac{H}{2} - L\right) + \frac{1}{2}\left(-\frac{H}{2}\right) = -\frac{L}{2}.$$

Because

$$\frac{H}{2} - L > \frac{L}{2} - L = -\frac{L}{2},$$

she does better to tell the truth.

Although such mechanisms have an artificial feel to them, they can often be converted into more realistic looking contracts. For instance, the above mechanism can be converted into an *option contract*: Party 1 is obligated to do the task and Party 2 is obligated to pay Party 1 $H/2$. However, Party 2 has the right (the option) to assume responsibility for the task if he wishes and, if he assumes responsibility, then Party 1 is obligated to pay him $L/2$. It is readily verified that Party 2 never gains from exercising his option if $c_2 = H$ and always gains if $c_2 = L$. Hence, Party 2 assumes responsibility only if he is low cost, which ensures that the task is always undertaken by the lower-cost party.

The above mechanisms are *balanced*, in the sense that any payment made by one of the parties is received by the other. The literature generally restricts itself to balanced mechanisms; in part, this is done because *un*balanced mechanisms seem at odds with what we observe in real life and, in part, because unbalanced mechanisms are too strong. The entire observable but unverifiable notion could be rendered completely meaningless, for instance, by using an unbalanced "shoot-them-both" mechanism: The judge asks the parties to simultaneously state what the observable variables are. If their reports agree, then the contract is enforced as required given the commonly reported realization of the variables. If their reports differ, then they are both shot (punished severely enough that neither side would willingly make a report that he or she knew differed from the other side's). Truth telling is an equilibrium of such a mechanism and, among the many equilibria, is arguably focal.[66]

4.3.2. The holdup problem and renegotiation

As discussed in 2.5.2 above, the possibility of holdup helps explain duress and related limitations on contractual freedom. The holdup problem has also been the subject of

[66] Although truth telling is a focal equilibrium given no pre-mechanism communication, it is not clear that this mechanism would work if pre-mechanism communication were permitted. That is, if you hear the other party say on the way up the court steps, "I will announce X," then you would also have to announce X even if that is neither the truth nor an advantageous statement.

intense research in the area of incomplete contracts.[67] This recent work has focused on the extent to which contract design can provide an alternative mechanism for policing holdup in situations where judicial intervention is unavailable or impractical. As such, this research can be seen as a particular application of the question, considered in the previous section, "does incompleteness matter?"

An example illustrates the problem. One firm, A, is to develop a product that a second firm, B, will market (A, *e.g.*, is a studio and B is a film distribution company). Because of its expertise, B can observe the quality of the product A has produced; yet quality could be sufficiently amorphous that it would be difficult, if not impossible, to describe in a contract or verify in court unambiguously. Setting incentives correctly would, thus, seem problematic. If B commits to buy the product at a fixed price, then A, acting opportunistically, has no incentive to invest enough in producing a quality product. If the contract is a revenue-sharing contract, then A and B's incentives could be too weak because of the teams problem (Holmstrom, 1982). Finally, if no contract is signed in advance, but the parties bargain after A has developed the product, then B could holdup A—that is, capture some of the value A creates by bargaining opportunistically—thereby adversely affecting A's incentives.

Demski and Sappington (1991) propose option contracts as a solution to this problem. Specifically, let $I \in \mathcal{I} \subset \mathbb{R}_+$ be A's investment (in dollars, say) and $\beta(I)$ be B's value of the product as a function of A's investment.[68] Assume for I and I' in $\mathcal{I}$ that if $I' > I$, then $\beta(I') \geq \beta(I)$. Suppose that I^* is the optimal investment for A to make. Demski and Sappington's solution is for the parties to sign an option contract whereby B has the right to acquire the product at a strike price of $\beta(I^*)$. If A underinvests, then B won't exercise the option, so A loses. It is a waste for A to overinvest. Hence, A should invest the appropriate amount exactly and B, being indifferent between exercising or letting the option lapse, can be assumed to exercise in equilibrium. Any desired sharing of expected surplus can be achieved through non-contingent transfers.

The Demski and Sappington solution is, unfortunately, vulnerable to renegotiation (Edlin and Hermalin, 2000). Recall that B is indifferent between exercising its option and letting it lapse, but exercises in equilibrium. Recognizing that, at the exercise date, its payoff is thus independent of its decision, B should pursue the following strategy: Let its option lapse, but then approach A about renegotiating a deal at a lower price.[69] Because there would otherwise be no rationale for trade, it must be that the product is

[67] A partial list of articles in this area is Bernheim and Whinston (1998), Che and Hausch (1999), Chung (1991), Demski and Sappington (1991), Edlin and Hermalin (2000), Edlin and Reichelstein (1996), Grossman and Hart (1986), Hermalin and Katz (1993), MacLeod and Malcolmson (1993), Nöldeke and Schmidt (1995, 1998), and Rogerson (1992).

[68] $\beta(I)$ could be an expected value and it subsumes B having made the optimal marketing decisions conditional on I.

[69] In a technical sense, this is *not* renegotiation insofar as the original contract has been honored—B had the right *not* to exercise the option. B is, in a technical sense, engaging in new negotiations following the expiration of the original contract.

worth less than $\beta(I^*)$ to A. Hence, there is a gain to the parties from negotiating a new price and, by the logic of the Coase theorem (see §2.2.2), they presumably will. Because the new price will be less than $\beta(I^*)$, this strategy of B's dominates a strategy of simply exercising the original option.[70] As a rational player, A will anticipate this; that is, that it will ultimately be held up despite the option contract. Anticipation of holdup will adversely affect A's incentives.

Some authors (*e.g.*, Bernheim and Whinston, 1998 and Nöldeke and Schmidt, 1998), have sought to restore the capacity of option contracts to solve the holdup problem. Their approach has been primarily to allow the parties to structure the option contracts in such a manner that the contracts are robust to the sort of renegotiation described above.

Edlin and Hermalin (2000) are less sanguine than these other authors about the parties' ability to make their contracts robust to renegotiation. Accordingly, they ask whether option contracts could still work, even given the threat of renegotiation, if the strike price is lowered from $\beta(I^*)$ to a more appropriate level? If the derivative at I^* of the product's value should it remain in A's hands is greater than $d\beta(I^*)/dI$, then it is possible to achieve the efficient outcome using an option contract (with a strike price below $\beta(I^*)$). If, however, $d\beta(I^*)/dI$ is greater, then Edlin and Hermalin prove that no contract (option or otherwise) can achieve the efficient outcome. To summarize: Recognizing the possibility of renegotiation, if any contract can achieve an efficient outcome, then an option contract can; but there are circumstances in which no contract achieves an efficient outcome.[71]

The literature on holdup problems yields, therefore, a mixed message with regard to the importance of the distinction between observable and verifiable. Under a variety of conditions, the distinction is ultimately meaningless, but there are circumstances in which the distinction is important. It is worth noting that, in terms of information, this literature has considered a rather stark world—if a variable is unverifiable, then it is (often implicitly) assumed that there is no correlated signal that is verifiable. As discussed in the previous subsection, were such signals to exist, then the circumstances in which this distinction mattered would be even further circumscribed.

4.4. *Legal doctrines addressing contractual incompleteness*

Given the complexity of the foregoing theoretical discussion, it should not be surprising that a significant portion of the law of contracts deals in one respect or another

[70] This argument doesn't hold if the bargaining game in renegotiation gives all the bargaining power to A; see Bernheim and Whinston (1998) and Edlin and Hermalin (2000) for differing views on the feasibility of (essentially) assigning the bargaining power to A *ex ante*. In addition, it could be welfare improving for the law to adopt rules that prevent renegotiation (Shavell, in press); although, as noted in footnote 69 *supra*, technically this bargaining between A and B is not renegotiation, but new negotiations following the expiration of the original contract.

[71] When first-best efficiency is not attainable, Edlin and Hermalin (2000) show that a simple sales contract is the second-best efficient contract. Because a simple sales contract can be replicated by an option contract with a very high strike price, there is a sense in which the optimal contract is always an option contract.

with problems of interpretation. Such problems include: how to determine whether the parties have entered into a contract, how to determine the terms of any contract that is formed, and what to do if the contract does not explicitly or implicitly cover the situation that has arisen *ex post*. It is not feasible to review this entire body of doctrine here (for such reviews, see the sources listed in §1.3.2), but we can survey the main regulatory strategies that the law uses to deal with the economic problem of incomplete contracts. We begin by discussing three major themes that pervade the law in this area—the role of default terms, the dichotomy of form and substance, and the tension between objective and subjective modes of interpretation—and then move to a series of specific doctrinal rules that illustrate these themes in operation.

4.4.1. Contract interpretation generally

The function of default rules. In §2.1 above, we discussed the role played by contractual default rules in regulating market failures. The primary function of such rules, however, is to provide guidance for those interpreting contract terms on which the parties do not otherwise clearly agree. For example, in an ordinary sale of goods, the default rule is that the buyer must pay in cash at the time of delivery; if the parties wish to provide for credit or some other medium of payment, they must specify so in their agreement. Such default rules extend to virtually all aspects of contractual agreements, including remedial terms such as damages; and no legal system could operate without them.

The law and economics literature has taken three main approaches to the question of how default terms should be set. These approaches should be understood as reflecting different means towards achieving the goal of an efficient system for completing incomplete contracts, each of which is premised on a different assumption regarding what types of transaction costs are most empirically common.

One approach, following the work of Goetz and Scott (1985), argues that default rules should be chosen to provide terms that would minimize the cumulative transaction costs incurred by parties contracting around them.[72] In the special case where all parties face similar contracting costs and it is equally costly to contract around all terms, for instance, this implies terms that would be favored by a majority. If, conversely, some terms are costlier to contract out of or into than others (for example, if it is harder for the parties to specify multi-factor standards than simple rules), then other things being equal the default should be set to the terms that are easiest for the parties to escape (Schwartz and Scott, 2004). One might call this the transaction-cost-reducing approach to default rules, with the caveat that the underlying concept of transaction costs may be ambiguously defined.

A second approach recognizes that contracting costs will lead many parties to stick with a default rule even if it is not the one they would have chosen. Hence, this approach

[72] This argument is typically presented in intuitive form, but it could easily be formalized along the lines of §2.2.2 *supra*.

advocates choosing default terms for their substantive efficiency. Such an approach is most appealing in settings where there are substantial network externalities, adverse selection, or bounded rationality in contracting, so that individual parties are reluctant to depart from familiar and widely-used terms. For instance, Korobkin (1998a, 1998b) presents evidence suggesting that cognitive and social-psychological factors lead parties to retain default (or standardized) contract terms that they would be better off contracting around; and Kahan and Klausner (1997) have argued that the network externalities are widespread in the corporate and business area due to the regular use of standardized forms, the value of which depends on a population of users who have invested in expertise in its application. This second approach might be called the central-command approach to default rules.

A third approach, following Ayres and Gertner (1989, 1992), argues that default rules should be used to encourage informationally advantaged parties to reveal their types before contracting, even if this means choosing terms that in equilibrium no one wants to use. They offer as an example the doctrine of *contra proferentem*, under which ambiguous contract terms are interpreted against the interest of the drafting party. Such an interpretation is unlikely to be what the drafter intended, they argue, but it serves as an incentive for the drafter to choose clear language that will convey to the counterparty the meaning of relevant terms. Ayres and Gertner, and much of the literature that follows, refer to such rules as penalty defaults, because they choose terms that operate as penalty for nondisclosure; an alternative and perhaps more precise terminology is to call this the information-forcing approach.

The concept of penalty default has been influential in the literature for its lucid illustration of the value of screening and signaling models in the analysis of contract law. Its general applicability in practice, however, is unclear from both positive and normative perspectives. From a normative standpoint, information forcing is only desirable under some circumstances; as we discuss in §3.3.3 and §5.2.7, it may be necessary to allow parties to keep the fruits of their informational advantage in order to induce them to acquire information initially. And even when it would be desirable to promote information revelation, a penalty default rule will fail to provide sufficient incentives in this regard if the value of the rents associated with informational advantage exceed the penalty. This can be the case if the informed party lacks bargaining power in a situation of bilateral monopoly (Johnston, 1990), if the information would reveal trade secrets valuable beyond the instant contract (Ben-Shahar and Bernstein, 2000), or if the asymmetric information is multidimensional and revelation along one dimension would allow the recipient to draw inferences about another (Adler, 1999). From a positive standpoint, conversely, E. Posner (2006) has recently argued that most actual legal rules do not fit the penalty default model, basing his argument both on a survey of Ayres and Gertner's original examples, and on case law generally.

Form versus substance in contract interpretation. The key policy question underlying contract interpretation is how thorough the interpretive process should be; and this question is commonly articulated in terms of the dichotomy of form and substance.

Specifically, many legal rules law have the effect of privileging certain interpretive materials and discounting others. Such rules are often termed formal or formalistic in that they confine attention to a subset of materials that may or may not give rise to the same inferences as would the universe of materials as a whole. A more substantive approach to contract interpretation, in contrast, would attempt to come to a more all-things-considered understanding, based on all of the materials reasonably available. For example, under the common law, the parol evidence rule (though subject to many exceptions) provides that a written document that integrates the parties' agreement may not be contradicted or varied by evidence of prior or contemporaneous oral understandings (E. Posner, 1998). In the absence of a writing, however, a party may introduce and a court can consider any information it wishes in order to determine the content of a contractual agreement.

From the economic viewpoint (Katz, 2004), the question of form versus substance can be assimilated to the problem of optimal information acquisition. A fuller or broader context can always be purchased, but at a cost of time, trouble, and interference with incentives, so it pays to stop at some optimal point. The problem can be formalized as follows: let the decision rule $D(\cdot)$ denote a function that maps from a set of facts to a legal outcome, for example, an award of damages. A court's *ex post* interpretation of the facts will depend upon the information available to it, which we can denote by information set $\mathcal{I}$. (For example, a court that is aware of commercial trade usage will interpret the words of a contract differently from one that is not.) This information set depends on three potential sources: the information embedded by the parties in the contract itself (denoted as $\mathcal{I}_0$), the information presented by the parties *ex post* at trial (denoted as $\mathcal{I}_1$), and the information available to the court based on its general knowledge and experience ($\mathcal{I}_a$). The information introduced by the parties in the contract or at trial is a variable that is either chosen in a cooperative game to maximize their returns from contracting or litigation, or the equilibrium outcome of a non-cooperative game in which they individually decide what information to introduce. Introducing information is costly, but the costs of *ex post* and *ex ante* information can interact, and more information can help the court reach a more efficient decision.

Within this framework, a legal interpretation regime consists of a function $R(\mathcal{I}_0, \mathcal{I}_1, \mathcal{I}_a)$ that specifies how the interpreting agent combines the three possible sources of information (for example, the most anti-formalist would be that the agent can consider all information available, in which case R would be the union function). A more formal regime would restrict R by limiting the influence of certain categories of information; for example, under the parol evidence rule, all information in $\mathcal{I}_1$ that was inconsistent with $\mathcal{I}_0$ would be thrown out. Conversely, a regime that disfavored standardized contracts on the theory that no one reads them would throw out any information in $\mathcal{I}_0$ that was inconsistent with $\mathcal{I}_1$. The framework is quite general and can incorporate many types of evidentiary systems; for instance, in a pure adversarial system, taking to an extreme the practices of common-law regimes, the agent could not take account of any information in $\mathcal{I}_a$ that was not confirmed by information introduced by the parties in $\mathcal{I}_1$. (Of course, the rule will affect the content of the information presented; if extrinsic information contradicting the written contract is ignored, no one will want to incur resources to present it.)

Given this framework, the expected legal outcome aggregating over all interpreting agents will be $V = \mathbb{E}_a\{D(R(\mathcal{I}_0, \mathcal{I}_1, \mathcal{I}_a))\}$. The parties will choose what information to embed in the contract, and to introduce at the time of a dispute, in order to maximize their individual returns in this outcome, and the lawmaker (who could be the parties themselves specifying in the contract what interpretation rule they wish to be followed) will choose the function R in order to maximize expected contractual surplus. From the standpoint of the lawmaker, this is just a constrained principal-agent problem, the solution of which depends on the particular costs associated with $\mathcal{I}_0$ and $\mathcal{I}_1$, and the incentive properties of the decision rule D with regard to the primary behavior governed by the contract (here suppressed in the notation, and assumed to occur at an intermediate stage between the writing of a contract and the potential emergence of a dispute).

The considerations that determine the optimal approach to contract interpretation are thus quite broad-ranging. The regime of contract interpretation will influence contracting parties' behavior in many respects: with regard to decisions to breach, to take advance precautions, to mitigate damages, to gather and communicate information, to allocate risk, to make reliance investments, to behave opportunistically, and to spend resources in litigation, and so on. Accordingly, it is difficult to draw strong general conclusions regarding how interpretation should proceed.

If one is willing to make restrictive assumptions about the costs and benefits of information acquisition, however, it is possible to reach more specific conclusions. For instance, Schwartz and Scott (2004) develop a model in which parties benefit when their contract terms are interpreted correctly on average, but are relatively indifferent (due to risk-neutrality) to interpretive variance around the mean. In this case it would be optimal for courts to make interpretations on a minimum evidentiary base, because additional interpretive efforts are costly but provide no incentive benefits from an *ex ante* perspective. Conversely, Shavell (2006a) assumes that *ex ante* contract-writing costs are positive and that it is possible to write at least some interpretive rules so that they can be applied costlessly (or more realistically, that the marginal cost of applying the rule is zero given that the parties are going to court anyway). In this case, it will be optimal for the parties to leave at least some open contract terms to be filled in *ex post* by courts, and if contract-writing costs are sufficiently high, for courts actually to override the written terms in favor of another interpretation.

Obviously, such assumptions are special; in their absence, perhaps the best that can be said is that private parties should be allowed the leeway to choose their favored interpretive regime—a leeway not always recognized by the legal system. For public lawmakers, who likely lack detailed contextual knowledge about the costs and benefits of information, however, it is possible to offer some general rules of thumb. For example, formal interpretation is more efficient, other things being equal, when (1) *ex ante* negotiation costs are low relative to renegotiation costs; (2) either the parties or the tribunal are likely to be biased in interpreting the contract; (3) the chance of a dispute (or the *ex post* stakes in the event of a dispute) are relatively high; (4) legal outcomes are relatively sensitive to litigation expenditure; (5) specific investments must be undertaken by persons distant from the transaction, or have value that is relatively context-independent; (6) the parties have relatively good control over their contract-drafting agents relative to their negotiating agents; and (7) the parties have relatively good access to non-legal sanctions. In all these situations, the expected costs of filling contractual gaps *ex ante* are low relative to the costs of filling them in *ex post*, so that it pays to undertake more effort in doing so up front. Conversely, when renegotiation costs are high, the extent of bias is significant, and so on, the expected costs of *ex post* gap-filling are relatively low and it pays to defer more of this effort until an interpretive dispute arises. The logic here is analogous to Kaplow's analyses of the choice between rules and standards (1992) and the optimal complexity of rules (1995).

It follows from these heuristic principles that substantive interpretation is relatively more valuable to small and infrequent traders, who are less well-placed to undertake the fixed cost of detailed *ex ante* negotiation; who have relatively poor access to reputational networks *ex post*; who are likely to do their own contract negotiating but to contract out when acquiring litigation services; who are less likely to be able to recover specific investments in substitute exchanges; and who possibly are less likely to face bias in *ex post* judicial tribunals. Conversely, large and experienced traders should prefer their contracts to be governed by relatively formalistic rules of interpretation. This distinction is consistent with casual empiricism—in general, it is experienced traders whom we observe contracting into relatively formal enforcement regimes through devices such as arbitration and forum selection clauses—and is also generally accepted in the case law and commentary.

Objective versus subjective interpretation. The default rule concept and the form-substance dichotomy both reflect the underlying problem of information acquisition and transmission, but focus on different dimensions of this problem. Specifically, the doctrines relating to form and substance are primarily concerned with the relationship between the contracting parties and the interpreting tribunal; they govern the extent to which the burden of information acquisition is allocated between the parties and the tribunal, and they affect the parties' incentives to transmit information about contractual meaning to the tribunal. Doctrines relating to default rules also implicate the relationship between parties and tribunal (especially when viewed from the transaction-cost-reducing approach of Goetz and Scott, 1985), but they additionally concern the contracting parties' informational relationship with each other (as emphasized by the penalty-default approach of Ayres and Gertner). These two dimensions of informational incentives interact in a third major theme of the law of interpretive doctrine: the tension between objective and subjective perspectives.

In this context, objective interpretation refers to the meaning that would be recognized by a reasonable outside observer, while subjective interpretation refers to the meaning as actually understood by the parties. For example, the case of *Lucy v. Zehmer* involved a dispute over whether the parties had agreed to the sale of the defendant's farm when, after a night of drinking, they signed a napkin that contained terms of sale, referred to the farm, and stated a price. The defendant claimed that he was joking when he signed the napkin, and that the plaintiff either knew or should have known this. The plaintiff for his part insisted that the deal was a serious one. In the end, the court sided with the plaintiff, but rather than emphasizing the plaintiff's state of mind as its justification, it referred to publicly observable factors such as the duration of the negotiations, the existence of previous negotiations between the parties, and the reasonableness of the contract price.

In general, however, official doctrine does not give clear indication of when objective meaning controls and when subjective meaning controls. The Second Restatement of Contracts (1981), generally recognized as the most influential modern summary of American doctrine in the area, states that subjective meaning controls if the parties attach the same meaning to an agreement, or if the parties attach different meanings to the agreement and one (and only one) of them is or should be aware of the other's understanding. (For instance, if the drafter of a form knows that the form contains an unusual term and also knows that the counterparty is unaware of the presence of the term, and the counterparty is not otherwise charged with knowledge of the term, then the term is not part of the contract.) On the other hand, more traditional authorities reject these principles and insist that unreasonable meanings will not be enforced; and the Restatement itself elsewhere provides that when the parties attach different meanings to the agreement and there is no asymmetric information between them, various objective factors will be used to determine the contract's meaning.

The practical complexity of the doctrine arises from the fact that it is being used to regulate information transmission both between the parties (*i.e.*, the problem of observability), and between the parties and the court (*i.e.*, the problem of verifiability). For example, the Restatement's rule that contracts are to be interpreted according to the meaning of the asymmetrically uninformed party plainly operates as a penalty default rule, and can be justified either as forcing disclosure or as protecting the distributional interests of the uninformed. And conversely, the rule that charges parties with knowledge of meanings of which they should be aware operates as an incentive for investigation.[73] But the objective-subjective distinction can also be understood in terms of the dichotomy of form and substance, in that an objective standard of interpretation directs the interpreter to limit its attention to factors that would be accessible to an objective observer, while a subjective interpretive standard directs the interpreter also to consider factors that might be accessible only to the parties to the contract. Thus, all the factors that bear on the optimal choice between formal and substantive interpretation (for example, the difficulty of litigating the parties' states of mind *ex post* as opposed to a counterparty's contrary meaning *ex ante*, or the likely variance of error when making such assessments) also bear on the choice between objective and subjective interpretation.

[73] Although, from a neoclassical economic perspective, it is unclear why it should be necessary to *provide* (strengthen) such an incentive. Presumably, private incentives should be sufficient (or even socially excessive) in this regard.

4.4.2. Specific interpretive doctrines

Parol evidence. In many cases, especially when substantial value is at stake, parties will memorialize their agreement in the form of a final written document. When they do, the intended effect with regard to pre-contractual communications can be ambiguous. For instance, suppose that, in oral negotiations, the buyer requests a clarification with respect to some technical feature of the goods and the seller provides clarification, but the final written document makes no mention of the issue. Should we regard the clarified term as implicitly incorporated into the final writing, or should we instead regard it as a rejected offer that the parties meant to exclude?

In common law systems, this question is addressed by a network of doctrines that are collectively referred to as the parol evidence rule. Under this rule, if the parties adopt a particular document as an authoritative statement of their contract—or in doctrinal language, as an integrated agreement—then the presumption is that they meant to reject all prior inconsistent communications. As a procedural matter, furthermore, the rule operates to exclude all evidence of antecedent understandings or negotiations that would vary or contradict the words of the document.

The parol evidence rule is subject to a number of exceptions and counter-doctrines (so much so that some commentators have suggested that it is misleading to refer to it as a rule); and the overall trend during the twentieth century has been toward decreased stringency in its application (typically, by finding, based on extrinsic evidence, that an apparently complete document was, despite superficial appearances, not intended by the parties to be authoritative). More recently, however, a number of commentators writing from an economically influenced perspective (*e.g.*, Scott, 2000 and Schwartz and Scott, 2004) have argued for its re-invigoration.

Formal economic analysis of the rule, however, has been limited. Eric Posner (1998) models the rule as posing a trade-off between Type 1 error (enforcing a term that is not part of the contract) and Type 2 error (failing to enforce a term that is part of the contract), as well as between transaction costs *ex ante* (*e.g.*, contract-writing costs) and *ex post* (*e.g.*, litigation costs). Posner's model should be understood as a special and more doctrinally detailed case of the general analysis of form and substance discussed *supra*. On his analysis, the rule makes little sense when courts process *ex post* information accurately and cheaply; it is justified when courts have poor ability to process *ex post* information and contract-writing costs are low. The implications for the more problematic situation where both contracting and litigation costs are high, however, are unclear and must presumably await future empirical investigation.

It is worth noting that the parol evidence rule does not apply to post-contractual communications, even though their use raises many of the same costs and benefits as pre-contractual communications do. Instead, such communications are analyzed using more flexible interpretive standards, and can operate as modifications or as waivers of contractual duty even when made informally.[74] It is possible that timing issues relating to renegotiation, holdup, and the increased value of *ex post* information could justify this disparate treatment, but at present this doctrinal distinction remains undertheorized from both legal and economic viewpoints.

Trade usage and course of dealing. In contrast to the disfavored category of parol evidence, other categories of material are given specially privileged status when interpreting individual contracts. The most important of these is trade usage, which refers to any practice of dealing that is sufficiently regular and widespread in the relevant area or line of trade as to justify an expectation that it will be observed in the particular case.[75] For example, a contract for the sale of two-by-four wooden planks will ordinarily be read to refer not to planks that are literally two inches by four inches, but to the smaller $1\frac{3}{4} \times 3\frac{1}{2}$ planks that generally pass under that description in the lumber business; a buyer who wants planks that are literally two by four must state so explicitly.

This doctrine can be justified in terms of both transaction cost reduction and information forcing. From the viewpoint of transaction cost reduction, if most contracting parties wish to follow a trade usage, it saves transactional resources for interpreters to read it into all contracts, so that only those who wish to disclaim it need undertake the extra costs of negotiating explicitly. From the viewpoint of information forcing, to the extent that one party subjectively intends to contract with regard to trade usage and the other does not, these differing understandings may lead them to enter into an inefficient exchange. Given that the trade usage is well evidenced and other usages are less so, requiring everyone to learn it offers a cheaper way of putting the parties on a similar informational footing.[76]

Such justifications are only persuasive, however, if trade usages can be identified cheaply and reliably, and if parties who wish to opt out of standard usage can do so at low cost. A number of recent commentators, however, have called both these assumptions into question. With regard to the first of these assumptions, Bernstein (1996, 1999) suggests that in many instances trade usages—in the sense of unwritten terms that are understood by the parties to impose actual legal obligations—do not actually exist, and that in any event, courts cannot tell the difference between behavior that evidences the existence of unwritten legal entitlements and behavior that instead represents settlement in the shadow of such entitlements. Also, Craswell (2000) has argued that courts often identify (or purport to identify) the actual content of any contested custom by consulting their own judgment, or the judgment of outside witnesses, about what custom would be most efficient. With regard to the second assumption, various commentators (*e.g.*, Scott,

[74] In some settings (*e.g.*, in sales contracts governed by Article 2 of the Uniform Commercial Code) it is possible for the parties to exclude oral modifications by using an appropriate clause in their original written contract, but it is not possible for them to exclude all possibility of *ex post* waiver.

[75] Restatement (Second) of Contract §222, Uniform Commercial Code §1-205.

[76] See Hermalin (2001) for a brief survey of research on the efficiency gains to be had from terminological conventions.

2002) have argued that although contracting around trade usage is permissible under official doctrine, courts in practice imply so heavy a presumption against it so as to make trade usage (or more properly, what courts consider as trade usage) effectively mandatory.

The related doctrine of course of dealing, which refers to a sequence of conduct between the parties to an individual contract, raises similar problems. Whether a pattern of bilateral behavior amounts to a legal understanding, however, is even more difficult to determine than with group behavior because such patterns tend to be less well evidenced and because the participants may wish to recharacterize their past behavior with an eye toward influencing the outcome of a current dispute. In addition, as Ben-Shahar (1999) has argued, allowing course of dealing to influence the interpretation of future contracts can lead parties to be excessively rigid when demanding contractual performance, for fear that greater flexibility will prejudice their future rights.

Such functional problems can be understood as stemming from the basic problem of contractual incompleteness, with the extra twist that it is not just the individual contract terms that are unverifiable, but also the alleged trade usage or course of dealing. Accordingly, whether courts should read trade usage and course of dealing into contracts as a default rule, or whether parties should make greater efforts to disclaim such a reading, remain open questions.

Change of circumstance. The doctrines of trade usage and course of dealing are used to fill out incomplete contracts in routine circumstances. In contrast, another set of doctrines are used to imply missing terms in unusual circumstances—specifically, when unforeseen events intervene to reduce substantially one or both parties' gains from trade under the original contract. For example, in *Taylor v. Caldwell*, a landlord whose building was destroyed in a fire was excused from liability to a promoter who had leased the premises for a theatrical event, even though the lease contract contained no explicit excuse of this sort; and in *Krell v. Henry*, a royalist socialite who had leased, at inflated short-term rates, a flat overlooking a public square in which the incoming British king was to be crowned was excused from the obligation to pay rent after the coronation was postponed due to the king's emergency appendectomy.

When this principle is used to excuse a producer or seller, it is often referred to as the doctrine of impossibility, although this term is a misnomer in that (1) the doctrine is often applied to situations in which it is highly burdensome but not impossible for the seller to perform, and (2) even if performance is not physically possible, the payment of damages is. Other labels include impracticability, commercial impracticability, and (especially when the buyer is excused, as in *Krell v. Henry*) frustration of purpose.

An analogous principle is sometimes used to excuse parties from complying with conditions that turn out to be unexpectedly burdensome, so long as the excuse does not unduly disadvantage the counterparty. (For example, a condition in an insurance contract requiring all claims to be made within two weeks of the occurrence of a casualty would typically be excused if the delay was caused by the insured party's being incapacitated in the underlying accident.) Finally, as indicated in §1.3.2 above, parties are sometimes excused from liability under the doctrine of mistake when they contract without the benefit of information that, once revealed, substantially alters the contract's efficiency or distribution of surplus. As observed above, from an information-theoretic perspective this situation is functionally equivalent to the case of *ex post* changed circumstances.

Economically influenced commentators have most commonly justified such doctrines on transaction-cost-reduction grounds. Specifically, Posner and Rosenfield (1977) suggest that, insofar as the doctrines apply only to excuse a party who neither has hidden information nor takes a hidden action with regard to the excusing event, they mimic the risk allocation terms that would most likely be provided in a complete contingent contract. (For instance, in *Taylor v. Caldwell*, given that the theater owner had already suffered the loss of a building, he was unlikely to be the least-cost insurer of the promoter's losses, at least in the absence of any specific facts evidencing moral hazard or adverse selection.)

The changed-circumstances doctrines can also be understood as special applications of the penalty-default and form-versus-substance frameworks outlined above. From a penalty default perspective, a contingency that could reasonably have been anticipated *ex ante* by one of the parties should not count as an excuse, because denying the excuse will encourage the informed party to bring the issue into the contractual negotiation. Conversely, from a form-substance perspective, contingencies that are easier to specify and ascertain *ex post*, either because information is better after the fact or because their probability is too low to warrant up-front attention, are appropriately left off to be dealt with by a court in the unlikely event that they arise.

It should be noted that the changed-circumstance doctrines can also be understood as special applications of the theory of optimal contract remedies along the lines of the discussion in §5 below. For instance, White (1988) points out that excusing a party for changed circumstances has consequences not just for risk allocation, but also for various countervailing considerations such as incentives for precaution and relational investment. She argues that it is only in rare cases that zero damages are second best, and accordingly that changed-circumstance cases should be treated like any other contractual breach. On the other hand, once the costs of legal enforcement are taken into account, excuse from liability may be warranted as a litigation-saving device to be applied in extreme situations where it is sufficiently clear that the contract should not be performed and that breach is not due to any failure of precaution or disclosure (in the same way that negligence liability in tort saves administrative costs by requiring litigation only in case of a breach of duty).

Standardized contracts. Most contracting parties do not negotiate individual contracts for each occasion on which they enter into exchange; instead they use standardized forms that incorporate customary or boilerplate terms, and negotiate only over those few terms (such as price and quantity) that are essential for their transactional purposes. From an economic viewpoint, such behavior is a straightforward response to scale economies in the production of contractual terms. In legal circles, however, standardized contracts are regarded with more ambivalence. While most legal commentators accept that such contracts are a practical necessity, many continue to regard them as problematic on grounds of autonomy (because most people do not read or understand other parties' standard forms and hence cannot knowingly assent to their terms) or distributive justice (because the party who drafts the form is deemed to hold an inequitable degree of bargaining power).

At an abstract level, standardized contracts are interpreted using the same methods as contracts more generally. For instance, if the drafting party knows that a particular term in the contract is unusual, that the non-drafting party is unaware of the term, and that the non-drafting party would not enter into the contract if aware of the term, then the term is not part of the contract; as observed in the discussion of objective versus subjective interpretation above, this result follows from ordinary principles regarding subjective interpretation. At a practical level, however, most lawyers understand form contracts to constitute a special category subject to specialized rules of interpretation, such as the doctrine that contracts are to be construed against the interests of the drafter. Conventional legal wisdom holds that form-contract terms are less likely to be enforced than individually negotiated terms, especially when they are unusual or thought to be especially burdensome to the non-drafting party.

There is relatively less reason to be concerned about standard terms on market power grounds, notwithstanding many legal commentators focus on this issue. Because standardization lowers the per-unit cost of contracting, both competitive firms and those with market power will find it of value. While firms with market power might distort non-price terms from the efficient level, the direction of distortion is *a priori* ambiguous (Spence, 1975). If all consumers have the same willingness to pay for contractual terms, a monopolist will do best to provide optimal terms and to extract available profits through a high price; and if consumers differ in their willingness to pay for contractual terms, super-optimal terms are as likely as suboptimal ones. While oligopolists could use standard forms to collude on non-price terms, it is similarly unclear in which direction their incentives cut. Accordingly, to the extent standard form contracts raise competition concerns, these are best policed by antitrust law, rather than by common-law courts applying the law of contract.

Standard form contracts do raise informational concerns, however, since they can vary substantially in their terms and the drafting party knows much more about the terms than does the non-drafting party. If contract reading is costly, parties may rationally assume that a particular contract contains the average terms available on the market, with the result being adverse selection along the lines of Akerlof's market for lemons (see also the discussion in §2.3.4). In addition, as Katz (1990b) has shown, if contract reading is a relation-specific investment (as it will be if different drafters offer different terms), drafters will be tempted to expropriate this investment by choosing terms that are just favorable enough for the non-drafter to accept after having read them, thus undermining any incentive to read. Interpretive rules that make it harder for drafters to enforce unusual standard terms without calling them to the other party's attention, such as the doctrine of *contra proferentem* or the rules requiring warranty disclaimers to be conspicuous, help to address this informational asymmetry by putting the burden of communication on the party who can undertake it most cheaply. Such rules provide perhaps the clearest illustration of the penalty default approach discussed above.

Implied duty of good faith. Finally, most modern-day courts will require the parties to a contract to exercise good faith in the performance and enforcement of their contractual duties. Commentators disagree on exactly what this obligation requires, but as a general matter it prohibits opportunistic actions that, while complying with the bare letter of the contract, depart from it in spirit by operating to deprive a counterparty of the reasonably expected benefit of its bargain. For example, it has been held a breach of good faith for a buyer of goods to seize on an otherwise trivial defect in delivery in order to escape an unfavorable contract in a falling market, for an employer to exercise its termination rights on an at-will labor contract just before the employee's accumulated sales commissions were due to be paid, or for a realty purchaser to prevent his seller-broker from acquiring the property to be delivered under the contract (and thus forcing the seller into breach) by showing up at an initial land auction and outbidding him.

In theory, the parties could guard against opportunistic exercises of discretion by providing specific limits in their original contract. For all the reasons discussed above in §4.2, however, they often do not do so, leaving courts with a disagreeable choice between imposing additional *ex post* limits on an *ad hoc* basis, and countenancing an unanticipated and perhaps unfair result. Traditional common-law courts often responded to this dilemma by refusing to enforce such agreements entirely, on the theory that such agreements were so one-sided as to lack formal consideration. Modern courts, in contrast, are more willing to treat such contracts as genuine bargains, and to police party opportunism directly and substantively after the fact. As our earlier discussion of form versus substance indicated, such an approach could serve the parties' economic interest if the courts' costs and error rate in filling contractual gaps *ex post* are sufficiently low, or if the parties' costs of dealing with them *ex ante* are sufficiently high.

Whether the duty of good faith is best viewed as a rule of interpretation, or alternatively as a limitation on contractual freedom, has not always been clear from the case law or commentary. The fact that the scope of the duty depends on other terms of the agreement, as well as on the overall commercial context, suggest that it is an interpretive rule; but the fact that the parties are not permitted to disclaim it (though they are permitted to stipulate its content so long as they do not do so in a "manifestly unreasonable" way) suggests that it is a regulatory intervention. The uncertainty is reinforced by the murky boundaries of the duty, which in limiting cases appears to shade into other mandatory doctrines such as unconscionability and public policy.

What counts as opportunistic behavior, however, typically lies in the eye of the beholder, and critics of the modern duty of good faith (*e.g.*, Schwartz, 1992; Bernstein, 1999; and Scott, 2000) argue that the behavior that it seeks to induce is typically unverifiable, and thus not effectively subject to judicial oversight. Courts' attempts to enforce good faith, accordingly, are as likely to depart from contractual expectations as to enforce them, in the end just resulting in an increase in the cost of litigation and in the variance of judicial outcomes. On this view, enforcement of such "soft" contractual expectations should be left to private enforcement mechanisms such as reputation and repeat dealing. For the most part, however, the courts have not yet accepted this critique.

4.5. Overall assessment of the law of contract interpretation

In general, the economics of contract interpretation is a relatively unmined field compared to the economic analysis of contract law generally. While the general contours of legal understanding in this area are more or less consonant with the main insights of the economic theory of contract, the rapid and recent development of the economic literature has not yet been matched by a corresponding growth in legal scholarship. It is likely, accordingly, that this set of topics will draw increasing attention from law and economics scholars in upcoming years.

5. Enforcement of contracts

As we observed at the outset of this chapter, the rationale for contracting is to lock in a commitment *ex ante* that one or both parties would otherwise not wish to honor *ex post*. The use of a contract to establish such commitment is undermined, of course, if the contract will not be enforced in the way the parties anticipate.

The enforcement of contracts is often the province of courts, who impose sanctions on parties who violate or breach a contractual obligation. As we discuss below, contracts can also be enforced by the parties themselves, through the use of various self-help remedies. We begin, however, with a discussion of judicial enforcement, partly because this has been the focus of most of the existing literature, and partly because an understanding of judicial enforcement is useful in order to understand when and why private enforcement might be more attractive.

5.1. General issues in enforcement

5.1.1. Remedies and contractual incompleteness

In one sense, every legal dispute is about remedies. Even when a court rules that no contract was ever formed, that ruling can be seen as saying that the complaining party is not entitled to any remedy for breach of contract. Moreover, such a ruling might still entitle the parties to some other legal remedy—for example, in some circumstances, one or both might be required to return any advance payments that had been made. From an economic standpoint, what ultimately matters is not the doctrinal question, "was a valid contract formed?" but rather the remedial question, "how much money can I now collect?" (or "how much money will I now be required to pay?").

In another sense, though, the identification of certain actions as a breach of contract, and the identification of certain payments as remedies, is an artifact of contractual incompleteness. After all, in a hypothetical complete contract, the contract itself would specify the payments and other actions that would be required in every possible state of the world. With such a contract in place, there would be no reason to label some actions in some states of the world as complying with the contract, while labeling other actions in other states as non-complying or breaching. There would also be no reason to label some (but not all) of the associated payoffs as "remedies for breach."

As discussed in §4.2 above, however, complete contingent contracts are rarely observed in the real world. In particular, real contracts are often incomplete in two ways that are relevant here. First, rather than specifying every possible action that a party might take, they may specify only those actions by each party that will suffice to comply with the contract—for example, "seller to deliver fifty widgets by July 1; buyer to pay $1000 by July 10." Second, they may fail to specify the steps that should be taken if either party fails to take a complying action—for example, if the seller delivers only forty widgets, or delivers them on July 15; or if the buyer refuses to pay.

Parties sometimes do provide explicit terms that specify the remedy for some forms of noncompliance; these terms are known as "liquidated damage clauses" and they will be discussed in §5.3.4. In other cases, however, where the parties' contract fails to specify the consequences of noncompliance, the legal system will supply a *default* remedy, just as it supplies default rules for other questions that a contract might not address (see §4.4). As most of the literature has focused on these default legal remedies, that is where we begin.

5.1.2. *Overview of default remedies*

In some cases, parties who fail to comply with their contracts may be ordered by courts to perform—in legal terms, specific performance (see §5.3.3 below), with the order backed up by threats of more severe sanctions for contempt of court. In most cases, though, the default remedy for breach has the breaching party pay money to the aggrieved party.

Following Fuller and Perdue (1936–37), it has become customary to identify the most common remedies as expectation damages, reliance damages, and restitution damages. *Expectation damages* attempt to put the injured party in as good a position as he would have been in if the contract had been performed. For example, if a buyer contracts to pay a price p for a good that has a gross value of v to him, but whose value can be realized only if the buyer spends r to make use of the good, full performance of the contract would leave the buyer with a net value of $v - r - p$. If the seller then breaches the contract, by failing to deliver the good, expectation damages would require the seller to pay enough to leave the buyer in that same net position: $v - r - p$. For example, if the buyer had already spent both p and r, expectation damages would require a payment of v.

By contrast, *reliance damages* require the seller to pay only enough to leave the buyer with the level of utility he would have enjoyed if the contract had not been signed. For example, if the buyer in this case would have had a net gain of zero had he not signed this contract, reliance damages would require the seller to pay enough to bring the buyer back to that position. So, for example, if the buyer had already spent both p and r, reliance damages would then equal $p + r$, enough to bring the buyer back to zero.

Another standard remedy, *restitution*, allows both parties to recover the reasonable value of whatever they gave to the other party. In this example, restitution would allow the buyer to recover damages of p.

As this example shows, these three remedies can often be ranked by size, with expectation damages usually providing the largest remedy and restitution giving the smallest. (We may assume $v > p + r$ because otherwise the buyer would not sign the contract.) Similarly, reliance usually exceeds restitution, because the buyer's reliance typically provides little benefit to the seller. Of these three remedies, expectation damages are conventionally regarded as the predominant measure in both theory and practice, with reliance and restitution reserved for cases in which there is some defect in the bargain or in the plaintiff's ability to prove it, or when equitable considerations justify a departure from the ordinary rule. In practice, however, this generality does not always hold, due to measurement difficulties and to other aspects of the legal rules governing the availability of each remedy.

For example, expectation damages entitle a plaintiff to compensation for profits that would have been earned on the breached contract, but calculating hypothetical revenues and expenses on a cancelled project is often a speculative endeavor. Accordingly, courts often presume that the future profits to be earned on the contract approximate the costs so far incurred, thus using reliance damages as a proxy for expectation damages. Conversely, reliance damages should in principle compensate a plaintiff for the loss of forgone alternatives, but such opportunity costs are often difficult to verify *ex post*. As a result, courts often presume that the forgone opportunity would have yielded the same profits as the breached contract—in which case expectation serves as a proxy for reliance. Similarly, it is often difficult to verify the value of restitution, especially when it takes the form of services not traded on any market. Thus, courts sometimes measure restitution in terms of the reliance costs incurred by the non-breacher, though their willingness to do so may depend on their perceptions of relative fault. There are even situations in which the non-breacher is permitted to choose among expectation and restitution, or among expectation and reliance; obviously, in those cases the more generous remedy will be chosen.

In addition, as we discuss below, each of these remedies can be modified or adjusted in various ways: for example, if the victim could have prevented some or all of the losses by appropriate mitigating behavior. As we also discuss below, there are a few additional remedies that are occasionally imposed following a contract breach, including punitive damages in rare cases.

In part for these reasons, it is often more helpful to think of a monetary damage remedy as a continuous choice variable d, whose value can in principle be set by the legal system to any level. In some cases, the damages awarded may more appropriately be written $d(\mathbf{x})$, where $\mathbf{x}$ is a vector representing any number of actions taken by one or both of the parties. Put less formally, remedies for breach can affect the parties' incentives both indirectly, just by threatening a party with having to pay some amount

d in certain states of the world; and more directly, by conditioning the size of d on some particular behavior x that the law seeks to influence. In general, though, damage awards that are conditioned on particular actions will often be more difficult for courts to implement, depending on the ease with which those actions can be verified by courts.

5.2. Monetary damages for breach of contract

We now turn to the effects of monetary remedies of various sizes and descriptions. Obviously, larger monetary remedies increase the non-breacher's payoff (and reduce the breacher's payoff) in the event of a breach, while smaller remedies produce the opposite effect. However, this effect by itself is merely distributional, and will not by itself change the transaction's expected value. As long as both parties correctly estimate the probability of a breach, the prospect of liability for higher (or lower) damages can be offset by charging a higher (or lower) price, leaving both parties with the same expected return.

Instead, changes in the expected value of the transaction must come from some other cause. For example, if one or both parties are risk averse, there can be welfare gains from shifting more of the variance in payoffs to the less risk-averse party. We discuss this possibility in §5.2.9 below.

More important, in most cases the expected value of the transaction depends in part on various actions the parties might take to increase (or reduce) that value. Moreover, the remedy for breach will usually affect the parties' incentive to take these actions—for example, their incentive to perform the contract rather than breaching, or to take precautions against contingencies that might lead them to breach, or to make transaction-specific investments whose value will be lost if there is a breach. Each of these incentive effects is discussed below.

5.2.1. The breacher's decision to perform or breach

Suppose that a seller must decide whether to perform a contract, at a cost c, when performance will confer on the buyer a value v. Total welfare is maximized if the seller performs when and only when $v \geq c$. Otherwise, it would be more efficient for the seller not to perform, an outcome often referred to as "efficient breach."[77]

To be sure, if v and c are known at the time of contracting, the parties would have no reason to contract unless v exceeded c. In many situations, however, c or v (or both) are

[77] The term "efficient breach," though widespread in the literature, is somewhat of a misnomer, and has had the unfortunate effect of inducing legal scholars untrained in economics to suppose that there is a tradeoff in this regard between efficiency and the deontological value of promise-keeping. From the perspective of a hypothetical complete contract, calling off performance when $v < c$ (possibly combined with some side payment) is precisely what the parties would have wished to specify, so a legal default rule that results in that outcome is better understood as promoting the parties' promissory intentions than subverting them. The term "efficient implied cancellation option" would be more accurate, if less arresting.

stochastic variables whose values will not be realized until just before performance is to take place, long after the contract was signed. Accordingly, one problem of interest is to design remedial rules that will give the seller an incentive to perform the contract if and only if $v \geq c$.

As the earliest contributions to this literature recognized, a remedy of expectation damages creates exactly this incentive, as long as those damages are accurately measured.[78] As defined above, expectation damages force the seller to internalize whatever losses the buyer suffers from the seller's breach. Consequently, the seller has an incentive to breach only if her gains from breach exceed both parties' losses. (A symmetric analysis, which we omit here for the sake of brevity, could be applied to the buyer's incentives to perform or breach.)

Before turning to other relevant incentives, four points should be made concerning even this simple result. First, while the expectation remedy forces the seller to internalize the buyer's losses from breach, it does not necessarily give the seller any incentive to consider effects that her performance or breach might have on third parties. As with most of the literature, we restrict attention to cases where third-party effects are absent (but see §2.3.1 and §5.3.4).

Second, like any other remedy, expectation damages will affect the seller's incentives only to the extent that the seller actually expects to pay those damages. If, instead, there is a significant chance that the seller could escape having to pay, the seller's incentives would be correspondingly reduced. As we discuss below (§5.4), the costs and other difficulties of pursuing a lawsuit may sometimes make this legal remedy less effective.

Third, if courts instead award some higher or lower measure of damages—or, equivalently, if they attempt to award expectation damages but predictably err in measuring those damages—the seller's incentives to breach will be altered. If courts tend to award damages that are higher than the true measure of expectation damages, the seller will have an even stronger incentive to avoid breaching: too strong an incentive, relative to the "efficient breach" condition described above. If courts instead tend to award lower damage measures, as is widely believed to be the case, the seller will have a weaker incentive to avoid breaching. The possibility of judicial error is particularly likely if some or all of the buyer's benefits from performance are either unobservable or nonverifiable (see §4.2 above).

Indeed, if courts instead could observe every relevant variable without error, it would be trivial to create efficient performance incentives simply by having the courts evaluate the efficiency of every breach *ex post*. If the court decided that a breach was efficient, the breacher could be excused from any remedy (or even rewarded with a bonus), while if the court decided a breach was inefficient, the breacher could be hit with huge sanctions. Viewed from this standpoint, the potential benefit of using expectation damages to create efficient incentives for performance or breach is that this remedy does not require courts to be able to evaluate the efficiency of any particular breach. The expectation remedy does, however, require courts to be able to calculate the amount the

[78] See Birmingham (1970) and Barton (1972). The first formal model is Shavell (1980).

buyer would have gained from performance. How well courts are able to make such calculations is a matter of some dispute.

Fourth, and most important, the incentives to perform or breach may not even matter as long as the parties can renegotiate after the values of v and c have been realized. If this *ex post* renegotiation is possible, then—regardless of the legal remedy for breach—there should always be an agreement that will maximize both parties' gains by performing the contract if but only if performance is efficient (Shavell, 1980). To be sure, renegotiation may entail positive transaction costs, but most legal remedies (including expectation damages) also entail positive transaction costs. While some of the early literature attempted to identify the legal remedy that would satisfy the "efficient breach" condition with the lowest possible transaction costs,[79] empirical evidence on actual transaction costs is largely nonexistent, and that literature is best described as inconclusive.

As discussed earlier, the possibility of *ex post* renegotiation also plays an important role in the literature on incomplete contracts (see §4.2.6 *supra*). This similarity should not be surprising—for as noted earlier—remedies for breach are really just one aspect of contractual incompleteness. As we discuss below, the possibility of *ex post* renegotiation also complicates the economic analysis of other incentives created by legal remedies.

5.2.2. The breacher's decision to take precautions

When the seller's costs c are stochastic, the distribution from which c is drawn might be given by nature, entirely independent of any action by the seller. More commonly, though, that distribution is itself a function of the seller's investment in the transaction. For example, a seller's costs might depend in part on how often her assembly line malfunctions; and the probability of such a malfunction might depend on how much the seller spent on maintenance. Typically, the seller must choose her expenditure on maintenance at one point in time; only then, after that expenditure has been made, the actual value of c will be realized.

Expenditures such as these are often referred to as "precautions" (see generally Cooter, 1985). The optimal expenditure on precautions can be defined straightforwardly in terms of (a) the cost of each possible precaution, and (b) its marginal effect on the distribution from which c is drawn, together with the welfare associated with each possible realization of c, taking account of the fact that some of those realizations will result in the contract being performed, while other realizations will result in the contract being breached. Given the usual assumption of diminishing marginal returns, there will be a unique expenditure on precautions that maximizes the total expected value of the transaction.

Interestingly, expectation damages (if measured accurately) can give sellers an incentive to choose this optimal expenditure on precautions (Kornhauser, 1983). As discussed above, expectation damages force the seller to internalize all of the buyer's losses from

[79] See, *e.g.*, Schwartz (1979), Macneil (1982), and Bishop (1985).

any realizations of c that result in a breach. As a consequence, expectation damages can in principle optimize both of the incentives discussed so far: (a) by giving the seller an incentive to choose the optimal expenditure on precautions; and (b) once the actual value of c is realized, by giving the seller an incentive to choose between performing and breaking the contract. Again, any remedies that are systematically less than expectation damages—whether by design, or because of measurement error—should reduce both of these incentives, while any remedies that systematically exceed expectation damages should increase both incentives.

To be sure, if courts could observe the seller's actual expenditures on precautions, and if they could also observe all the factors necessary to calculate the optimal expenditure, there would then be other ways to give sellers an incentive to choose the optimal level. For example, courts could excuse from liability any seller whose precautions were optimal, in much the same way that negligence standards in tort law excuse defendants whose precautions were optimal. Alternatively, if the contracting parties could specify in their contract the optimal expenditure on precautions, that requirement could itself be enforced by a court, as long as the court could verify the seller's actual expenditure. Thus, just as in the case of incentives for performance (§5.2.1), the case for using expectation damages to optimize a seller's incentives to take precautions rests in part on the assumption that it is easier for courts to observe only those factors necessary for the calculation of expectation damages (*i.e.*, only those factors that go to the value that the buyer would have received if the contract had been performed) than it is for courts to observe any of these other factors.[80]

Alternatively, if buyers can observe a seller's precautions (before contracting), competition among sellers could give them a market incentive to choose the optimal expenditure on precautions (Kornhauser, 1983). Even if sellers do not choose their precautions until after contracting, optimal market incentives could be possible if buyers can observe sellers' reputation for taking precaution. In §5.4 below, we consider reputations and competitive markets in connection with enforcement by means other than legal remedies.

5.2.3. The non-breacher's reliance decision

The value the buyer places on performance by the seller is often a function of the buyer's investment in the transaction. For example, a business that is buying a new machine may get more value from that machine if it spends money training its employees to use it. Expenditures such as these are often referred to as reliance on the contract.

By definition, reliance expenditures are at least partially transaction-specific—for example, training designed for one particular machine may be worthless if that machine is not delivered. Although there can often be varying degrees of performance, for concreteness assume that there are only two discrete possibilities—either the contract is

[80] For a discussion of this aspect of expectation damages, see Cooter (1985).

performed, or it is breached. Let $v(r)$ represent the value that the buyer will receive from performance conditional on his reliance investment r, while $w(r)$ represents his value if the contract is not performed. If performance of the contract is certain, the optimal reliance expenditure is simply the value that maximizes $v(r) - r$. In the more usual case, where the contract should be performed in some states of the world only, the optimal reliance expenditure is the value that maximizes

$$qv(r) + (1 - q)w(r) - r, \tag{9}$$

where q is the probability that contract should be performed (here taken to be independent of r; *e.g.*, with probability $1 - q$ some "act of God" makes performance prohibitively expensive for the seller).

The remedy of expectation damages ($d = v(r) - w(r) - p$, in this example) will generally not give the buyer the right incentive to choose the optimal value of r. Expectation damages are a "full insurance" remedy that gives the buyer the same net return in every state of the world; that is, the buyer's net return is $v(r) - p$ under performance and $d + w(r) = v(r) - p$ under non-performance. Unless $v'(r^*) = w'(r^*)$, where r^* maximizes expression (9), the buyer will choose a non-optimal level of reliance. If one assumes, as is often true, that the marginal return from reliance investment is greater given performance than non-performance (*i.e.*, $v'(r) > w'(r)$), then buyers have excessive incentives to rely under expectation damages.

Reliance damages ($d = r - w(r)$, in this example) also will fail to provide the buyer the proper incentives. If $v'(r) > w'(r)$, then $w'(r^*) < 1$; in this case, reliance damages will encourage the buyer to rely excessively because increased expenditures on r will increase the damages they can collect.

More generally, the probability q is endogenous insofar as the level of damages (either expectation or reliance) affects the probability that the contract will in fact be performed (see §5.2.1 and §5.2.2). Even with an endogenous q, however, neither reliance nor expectation damages will generally provide the proper incentives. To see this, observe that the first-best solution requires performance whenever $v(r) - c \geq w(r)$ and, therefore, that investments maximize

$$\int_0^{v(r)-w(r)} (v(r) - c) f(c|s) dc + \int_{v(r)-w(r)}^{\infty} w(r) f(c|s) dc - r - s, \tag{10}$$

where s is the seller's investment in precaution and $f(c|s)$ is the density of cost given such investment. To replicate the first-best solution with regard to whether the seller performs, it must be that $p - c \geq -d$ if and only if $v(r) - c \geq w(r)$; hence, $d = v(r) - w(r) - p$, which accords with expectation damages, but not reliance damages. However, as we saw above, expectation damages causes the buyer to face the maximization problem $v(r) - r$ with respect to his choice of reliance. As discussed, the r that solves the buyer's problem will differ from the r that maximizes social surplus (*i.e.*, expression (10)).[81]

[81] See Shavell (1980) for greater details. Consider also Shavell (1984), Kornhauser (1983), and Rogerson (1984).

If and only if *ex post* renegotiation is impossible, a remedy of no damages at all—*i.e.*, a regime of no liability for breach—can optimize the buyer's reliance incentives in a second-best sense (given the probability of performance), because such a regime forces the buyer to bear all of the downside as well as all of the upside of any reliance expenditure (Shavell, 1980). To be sure, such a remedy does little to optimize the seller's incentives, as discussed earlier in §5.2.1 and §5.2.2. But if the breach probability (*i.e.*, q above) is exogenous or the seller's incentives can be optimized independently (*e.g.*, through market-reputation effects), a zero-damage remedy could be optimal. This is not wholly surprising, as this forces the buyer to face the social planner's problem as given by (9) above.

Mathematically, zero is just a constant and, more generally, any constant-damage measure—that is, any measure whose value does not change with the buyer's level of reliance—will also optimize the buyer's reliance incentives (for any given probability of performance), as long as *ex post* renegotiation is still impossible.[82] Under a constant-damage measure, the buyer will still capture all of the marginal benefits, and bear all of the marginal costs, of higher or lower reliance expenditures. Moreover, a constant-damage measure could also optimize the seller's incentives, if the constant is set equal to the value that performance would have to a buyer who in fact relied optimally: $d = v(r^*) - w(r^*) - p$, in this example. In addition to optimizing the buyer's reliance expenditures, this will also give the seller optimal incentives, both to perform or breach and to choose an optimal level of precautions.[83]

However, as noted, this result holds only when *ex post* renegotiation is impossible. When renegotiation is possible, reliance by the buyer can have the additional effect of altering the terms likely to be reached in any renegotiation, thus further distorting the buyer's reliance incentives (recall the discussion of holdup above in §4.3.2). Under certain conditions, holdup concerns may bias the buyer's incentives toward choosing too little reliance (Rogerson, 1984).

Edlin and Reichelstein (1996) show that, in some circumstances, the parties may be able to choose contract terms that balance the risk of holdup against the risk of moral hazard, thus implementing a first-best solution (recall, too, the discussion in §4.3.2). For example, suppose the parties enter into a fixed-price contract in which they can freely adjust the quantity term, and again, suppose that the buyer must rely before he knows whether the seller will perform. If the contract quantity is set at zero (*i.e.*, if there is no enforceable contract at all), the buyer will under-rely because part of the value created by reliance will be expropriated through *ex post* holdup. On the other hand, if the contract quantity is set at the level Q^* that the buyer wishes to purchase, he will be fully insured against non-performance, and will over-rely. By continuity, there exists some contract quantity $\hat{Q} \in (0, Q^*)$ that will provide optimal reliance incentives. In the event that the seller can perform, the parties can then make up the difference between $\hat{Q}$ and Q^* *ex post*, entering into a spot transaction for an additional amount $Q^* - \hat{Q}$.

[82] Shavell (1980) considers a model in which restitution is a constant-damage measure.

[83] See Cooter (1985) and Cooter and Eisenberg (1985).

Alternatively, buyers' reliance incentives can always be optimized if courts are capable of evaluating directly the efficiency of any reliance expenditure. If courts can evaluate reliance expenditures directly, then it is easy to create optimal incentives simply by rewarding any buyer that relies optimally or by penalizing any buyer who fails to rely optimally. Indeed, there are various legal doctrines that might be interpreted as having this effect by, for example, awarding damages only for "reasonable" reliance expenditures (Goetz and Scott, 1980), or by calculating recovery based on the reliance that would have been "reasonably foreseeable" (Cooter, 1985). Obviously, the effect of any such mechanisms depends on how courts define a "reasonable" (or "reasonably foreseeable") level of reliance. This, in turn, depends heavily on the verifiability of the factors that define the optimal level of reliance (the $v(\cdot)$ and $w(\cdot)$ functions, in the model sketched here).

Of course, if the parties themselves can write a complete contract, reasonable expenditures on reliance could be spelled out in the contract itself. As noted earlier, though, this much contractual completeness is sometimes difficult or impossible. Alternatively, if parties can acquire reputations for relying optimally (or for relying excessively), then the competitive advantages of such a reputation could also induce the parties to make optimal reliance investments. We defer discussion of these possibilities to the section on non-legal enforcement (§5.4).

5.2.4. *The non-breacher's precautions*

A closely-related issue concerns precautions that the buyer may be able to take to reduce the loss he will suffer if the seller breaches. For example, if there is a significant chance that the seller may not deliver the machines she promised, it might be efficient for the buyer to keep his old machines as back-ups rather than getting rid of them as soon as the contract is signed, even if keeping the old machines around is costly. If the seller does perform the contract, the expenses involved in keeping the old machines will have been unnecessary—but if the seller fails to perform, those expenses will have been well spent. Conceptually, then, *failing* to take such a precaution can be thought of as another form of reliance on the promise of performance (Cooter, 1985).

As a consequence, the legal remedy for breach can also affect buyers' incentives to take these sort of precautions. As discussed above, simple expectation or reliance remedies create a moral hazard problem that can leave buyers with incentives to rely too heavily—or, in this context, to take too few precautions. There are, however, legal doctrines that might correct that problem by limiting buyers' ability to recover for losses that could or should have been avoided by appropriate pre-breach precautions. The doctrine of *Hadley v. Baxendale*, for example, may sometimes be used to deny recovery of losses that the buyer could have prevented, by ruling that such losses were not "reasonably foreseeable" to the seller (*e.g.*, R. Posner, 2003, p. 127). Similarly, the implied excuse doctrines—impracticability, frustration, and mistake—could in principle be used to optimize buyer's incentives by releasing the seller from liability entirely, thereby making the buyer bear all losses suffered from the seller's nonperformance in just those cases where the buyer should have taken more precautions than he did (Posner and Rosenfield, 1977). If courts release sellers from liability only when buyers have behaved suboptimally, in a manner similar to a contributory negligence rule in tort law, that could give buyers an incentive to optimize their own behavior. And as long as buyers optimize their own behavior, courts could then return to holding sellers fully liable, thereby (in principle) optimizing both parties' incentives.

However, these doctrines can have this effect only if courts are able to evaluate fully the precautions available to buyers *ex ante*, in order to recognize cases when the buyer should indeed have taken additional precautions. The extent of courts' ability to do this—and to do it reliably enough so that buyers will know when they will be held liable if they do not take precautions—is a matter of some dispute.[84]

5.2.5. The non-breacher's mitigation of losses

Another common form of precaution involves steps the buyer takes after the seller's breach becomes final. For example, if the buyer has bought special equipment to use with a machine that is never delivered, that equipment can sometimes be salvaged or reused, though perhaps not for its full original value. In the notation used in the preceding section, this salvage value was implicit in the $w(r)$ function, but that notation suppresses the fact that the buyer (the non-breacher) may have to take various steps in order to realize that salvage value. Moreover, some such steps are likely to be efficient while others are not—for example, in some cases the cost that must be incurred to salvage unused equipment may exceed the equipment's salvage value.

Courts are rather more willing to consider such *ex post* opportunities for mitigation of losses and to require buyers to take advantage of them than they are to consider the precautions that buyers might have taken before breach. In part this is because the value of mitigation is especially salient after the fact of breach (or from an incomplete contracts perspective, more easily verifiable). As a matter of law, this issue is most explicitly addressed by the mitigation doctrine, which limits buyers' remedies by denying compensation for any losses that could have been avoided by reasonable mitigation.

As with many legal concepts, what counts as "reasonable" mitigation is a matter of some dispute. To the extent that courts' definition of reasonable mitigation corresponds with efficient mitigation, this limit on remedies can give buyers efficient incentives along this dimension.[85] Obviously, though, the efficacy of this mechanism depends on courts' ability to verify the costs and benefits of various mitigation activities.

5.2.6. The decision to terminate a project

In many cases, performance of a contract requires a sequence of many choices or events. For example, the construction of a building involves hundreds of separate steps whose

[84] For a skeptical view, see Kull (1991).

[85] For discussions of this possibility, see Wittman (1981) and Goetz and Scott (1983).

performance can extend over weeks or months. Moreover, sometimes the earliest steps in the sequence must be taken at a time when it is still uncertain whether it will be worthwhile to perform all of the later steps. For example, if future construction costs are uncertain, it might (or might not) prove uneconomic to finish the building, but the early stages of construction may have to start before anyone can be certain what the eventual costs will be.

In such cases, each step in the sequence can be thought of as involving a choice between (a) continuing to perform, or (b) giving up the attempt to perform by terminating the project. Terminating the project early can allow the parties to take steps to reduce their losses, as discussed in the previous section on mitigation. On the other hand, continuing to perform preserves the option value of the project, by deferring to each later stage the decision to continue or terminate. The optimal decision, therefore, is to terminate early only when the savings from doing so exceed the project's option value (Triantis and Triantis, 1998).

The doctrine of anticipatory repudiation allows the potential breacher—the seller, in the examples above—to terminate a contract early simply by repudiating it.[86] While this renders the seller liable for damages for breach, it also triggers the mitigation doctrine discussed above, thus requiring the buyer to begin taking any available steps to reduce those damages. As a result, if the seller faces a high probability of being unable to perform, she may be able to reduce her eventual liability by terminating the contract early. Viewed from this perspective, a termination decision by the seller can be thought of as just another form of breach, which will be efficient (or not) depending on whether early termination is optimal.

Indeed, if the seller can be held fully liable for all the losses from breach—either all the current losses, if the contract is terminated early, or all the subsequent losses if the contract is not terminated early—then she will have an incentive to make the optimal choice between continuing or terminating early, for much the same reasons discussed in the section on efficient breach (§5.2.1). However, holding the repudiating party liable for all losses may be difficult. The losses from an early termination include the loss of the option value of the contract, which may be difficult for courts to measure; omitting this loss from the measure of damages will bias the seller's incentives toward terminating too early (Triantis and Triantis, 1998). At the same time, the losses from any eventual non-performance may also be difficult to measure; or they may be difficult to collect, if the seller is by then insolvent. Omitting these losses from the damage measure will bias the seller toward terminating too late (Craswell, 1990).

Perhaps because of these difficulties, various legal doctrines also allow the non-breaching party—the buyer, in the example used here—to force an early termination (while still holding the other party liable for damages) subject to *ex post* judicial review. From the buyer's standpoint, a decision to terminate early can be thought of as another

[86] For discussions of the measurement of damages in cases involving anticipatory repudiation, see Jackson (1978) and Goetz and Scott (1983).

form of precaution (albeit an extreme one), which reduces the losses that will be suffered in the event of a breach. Equivalently, a decision not to terminate early can be thought of as another form of reliance by the buyer, since such a decision increases the gains if the project is eventually successful, but also increases the losses if the project eventually fails. In keeping with the earlier analysis of reliance and precautions (§5.2.3 and §5.2.4), the buyer will not have an incentive to make this decision optimally if he is fully insured by being guaranteed full compensation for his losses.

Perhaps for this reason, the law does not give the buyer the unrestrained power to elect an early termination of the contract, but (using various legal doctrines) subjects that decision to judicial review. For example, if the seller has already breached the contract during early stages of performance, the buyer can elect to force an early termination as long as those breaches are judged by a court to be "material" or "total." Alternatively, even if the seller has not yet violated any contractual requirement, the buyer may still elect to force an early termination if he has "reasonable grounds for insecurity" about the seller's performance, and if the seller is unable to provide "adequate assurances" that she will be able to perform.

Of course, the exact effects of these doctrines depend on how courts interpret such vague standards as "material," "reasonable," or "adequate." (If courts could observe all the factors necessary to determine whether early termination was optimal, they could allow termination in exactly those cases; but their ability to do this is doubtful.) The issue is further complicated by the fact that courts do not usually issue advance opinions on pending disputes, but instead render their judgments in hindsight. An insecure buyer must thus decide whether to force an early termination before he is certain whether a court will agree that the other party's earlier breach was "material," or that the other party's assurances were not "adequate." If the buyer tries to terminate early, and a court later rules that grounds for early termination were lacking, the buyer may himself be held liable for breaching the contract prematurely. Given the substantial uncertainty associated with such vague legal standards, a risk-averse buyer may prefer not to force the issue (and a less risk-averse seller may take advantage of the situation to force a modification or a release of its obligation to perform).

Finally, as with other perform-or-breach decisions, the decision to terminate early should be made optimally whenever *ex post* renegotiation is possible, as there should always be some combination of side-payments that makes it in both parties' interests to terminate early whenever early termination is optimal. However, the legal doctrines discussed above will affect each party's share of any surplus, depending (for instance) on whether the buyer has a unilateral right to force an early termination without the seller's consent. As usual, though, such renegotiation may itself be costly; and the prospect of higher or lower side-payments (in the event of renegotiation) can also alter many of the other incentives discussed above, such as the incentive to rely or the incentive to take precautions.[87]

[87] For further discussion of all of these issues, see Goetz and Scott (1983), Craswell (1990), and Triantis and Triantis (1998).

5.2.7. Decisions to gather and disclose information

Much of the literature described in the preceding subsections presupposes that both parties already know all of the relevant parameters, such as the distribution from which the seller's cost, c, will be drawn, or the extent of a buyer's reliance investments, r. In many cases, though, this (and other) information cannot be discovered without costly investments in information-gathering. To be sure, a number of legal doctrines affect the incentives to gather or disclose information, including some that are not normally classified as "remedies" doctrines (for example, see the discussions of fraud and mistake earlier in §2.5.2 and the duty to disclose in §3.3.3). However, the remedies for breach can also affect various information-related incentives. While there are many different kinds of information that might be gathered or disclosed, and the effects of remedies on these incentives have not been studied as extensively as have the other incentives discussed earlier, we discuss some representative examples.

Deciding whether to enter into contracts. The greater the penalties for breach, the greater will be sellers' incentive to gather information about potential risks that might leave them unable to perform. If a seller who fails to perform is held fully liable for all of a buyer's losses, that will make her internalize the full costs of any failure on her part to gather sufficient information. However, this will not be sufficient to make sellers' incentives optimal unless sellers can also internalize the full benefits of any additional investments in information-gathering. A perfectly price-discriminating monopolist may be able to internalize all of those benefits, by charging higher prices; and in perfectly competitive markets, competitive pressures may force sellers to account for those benefits as well. In markets that fall between these extremes, however, it is more difficult to design a remedy that gives sellers the optimal incentive to gather this information (Craswell, 1988).

Searching for contractual partners. In many markets sellers (and buyers) may have a choice as to how much effort they devote to finding potential trading partners. Here, too, the parties' incentives depend partly on other legal doctrines—in this case, the rules governing initial contract formation (see §3 above). However, the parties' incentives on this point, too, will also depend on the general remedies for breach. For example, the steeper the penalties for breach, the harder it will be to get out of any contract once a contract has been formed, and so the more it will pay parties to invest in searching longer and harder to make sure they have found the best deal. Here, too, some of the benefits of increased search may be felt by the party who is found, rather than by the party who is doing the searching, so efficient search incentives can be created only if the searching party is able to internalize these benefits. For formal models incorporating these effects, see Diamond and Maskin (1979, 1981).

Optimal precaution and reliance. Even after one party has gathered information, in some cases the other party is the one who needs to be given that information in order to make an optimal decision. For example, if the efficient level of seller precautions depends on how much the buyer would gain from performance, and if the buyer already knows that information, the seller's precaution incentives can more easily be optimized if the buyer can communicate his information to the seller (Bebchuk and Shavell, 1991). Similarly, if the optimal level of reliance by the buyer depends in part on the probability that the seller will fail to perform, and if the seller already knows that probability, the buyer may be better able to choose an optimal level of reliance if the seller tells him what that probability is (Craswell, 1989b). While most of the conventional remedies for breach do not create any incentives for a buyer to convey this information to the seller, some remedies affect this incentive by conditioning the measure of damages on the information that has been disclosed. Under the rule of *Hadley v. Baxendale*, for example, a buyer facing large losses from breach is more likely to be allowed to recover those losses if she has told the seller about them in advance.[88]

5.2.8. The effects of party heterogeneity

Information also matters if sellers (or buyers) must deal with a heterogeneous population of trading partners. In many settings, for example, sellers may differ in the probability that they will be able to perform, or buyers may differ in the amount they would lose if the contract were breached. If each party's type is fully known to the other party, those differences can be fully reflected in each side's prices and other behavior, allowing each interaction to be analyzed as if it involved entirely homogeneous parties. When differences in characteristics cannot be fully observed, however, much of the analysis described above must be altered in some way.

For example, if buyers differ in the amount they will lose from a breach—and if the law's damage measure reflects these differences, by holding sellers liable for different damages depending on the buyer's type—sellers' expected liability costs will be lower when dealing with low-damage buyers than when dealing with high-damage buyers. If sellers can recognize the high-risk buyers, they can adjust their prices or their level of precautions or both to reflect those greater risks. However, this gives high-damage buyers an incentive to conceal their greater riskiness, if they can. In that case, the equilibrium will depend on the ability of the low-damage buyer to signal his type or the seller's ability to screen buyer types. As discussed in §2.3.2, signaling and screening can result in a loss of welfare. If screening or signaling are ineffective at separating the types, then a pooling equilibrium could result in which low-damage buyers effectively subsidize high-damage buyers. Alternatively, to avoid this subsidy, low-damage buyers could exit; that is, an adverse selection ("lemons") problem ensues. Limits on

[88] For economic analyses of this aspect of the Hadley doctrine see, *e.g.*, Ayres and Gertner (1989), Wolcher (1989), Johnston (1990), Bebchuk and Shavell (1991), Adler (1999), and Ben-Shahar and Bernstein (2000). Because of possible adverse price discrimination *vis-à-vis* the shipping price, however, a buyer may be reluctant to reveal that she faces large losses (see §4.2.4 and §5.2.8).

the recovery of unusually high damages—including the doctrine of *Hadley v. Baxendale*, discussed in the preceding subsection—could conceivably be used as a screen to prevent subsidization or drop out (see Quillen, 1988).

Buyer heterogeneity is also important if a seller has market power, insofar as it introduces the possibility of price discrimination. To be sure, if the seller can fully observe each buyer's type, she can engage in first-degree price discrimination, a practice that general contract law does not restrict. But if the seller cannot observe each buyer's type directly, she may be able to use some of the terms of her contract (including the remedies for breach) to separate different classes of buyers—screen—thereby increasing her profits through second-degree price discrimination (see, *e.g.*, Matthews and Moore, 1987). This possibility could also be relevant to the analysis of liquidated damage clauses, discussed later in §5.3.4. As discussed in §2.3.3, especially footnote 29, the welfare consequences of such price discrimination are *a priori* ambiguous.

5.2.9. *Risk allocation and insurance*

Finally, in addition to all of the incentive effects discussed above, remedies for breach also have the effect of allocating various risks between the two parties. For example, the expectation remedy defined in §5.1.2 will, if it is measured perfectly, leave the buyer (the non-breaching party) fully insured. Thus, if buyers are risk averse, while sellers are risk neutral, this remedy will be efficient in terms of its affects on the parties' attitudes toward risk. For other combinations of attitudes toward risk, remedies other than expectation damages will be superior (see generally Polinsky, 1983). Unless one of the parties is actually risk loving, however, remedies that exceed the non-breacher's actual losses (*e.g.*, punitive damages) will almost always be undesirable as far as risk sharing is concerned (Craswell, 1996a).[89]

One case of interest concerns breaches that inflict non-pecuniary losses that do not increase the buyer's marginal utility of money (see Cook and Graham, 1977). For example, a photographer's failure to take wedding pictures might reduce the welfare of the marrying couple, but that does not mean they would prefer to buy an insurance policy that would pay them additional money if their wedding pictures were lost.[90] Viewed purely from the standpoint of the parties' taste for insurance, then, it might be better if contract remedies did not compensate for this sort of loss (as, indeed, they generally do not). At the same time, though, excluding these losses from contract remedies could also distort some of the other incentives discussed above, such as the photographer's incentive to take adequate precautions. For an analysis incorporating both of these effects, see Rea (1982).

[89] Risk aversion also plays a role in why limits on damages could be Pareto improving when the damage terms signal information about the potential breacher. See §2.3.2 and the discussion of Aghion and Hermalin (1990).

[90] Such preferences are, admittedly, inconsistent with neoclassical economic theory unless the couple exhibits lexicographic preferences in photos-money (other goods) space.

5.3. Complications in determining monetary damages

In the discussion so far, we have dealt with the incentive and insurance effects induced by monetary remedies for breach. To be sure, much of the literature on this topic developed by analyzing particular legal remedies: either particular measures of monetary recovery, or non-monetary remedies such as specific performance. Reflecting that literature, §5.3.1 considers various legal and practical limits on contract damages; and section §5.3.2 considers some alternative damage measures that are imposed in certain cases. Then, §5.3.3 and §5.3.4 discuss specific performance (injunctive relief) and liquidated damage clauses (remedies stipulated in the contract itself).

5.3.1. Limits on the measure of damages

In theory, expectation damages leaves the non-breacher just as well off in every state of the world (see §5.1.2); but in practice, a number of legal doctrines limit the losses that expectation damages will compensate. For instance, the non-breacher must prove the amount of his loss with "reasonable certainty"; often this will exclude the recovery of "speculative" losses whose amount was uncertain. Also, as noted earlier, contract law only rarely allows compensation for emotional losses (see §5.2.9). Perhaps more significant in this regard than any rule of contract law, however, is the general default rule of US legal procedure that requires each party to bear his own costs in litigation. In most other Western legal systems, in contrast, prevailing litigants are entitled to recover attorneys' fees as well as other out-of-pocket litigation costs. As a result, they come closer to being made whole than do winning litigants in the US.

By reducing the effective amount of the remedy, doctrines such as these weaken many of the seller's incentives discussed earlier (for example, the seller's incentive to take precautions, see §5.2.2). Of course, by shifting more of the loss to the buyer, the same doctrines may also strengthen the buyer's incentive to take various precautions (see, *e.g.*, §5.2.4). Also, when these exclusions are conditioned on particular behavior by the buyer—*e.g.*, if recovery for certain losses is excluded unless the possibility of such losses is disclosed in advance, under the doctrine of *Hadley v. Baxendale*—that could strengthen the buyer's incentives in more focused ways (see §5.2.7). Finally, if buyers differ in the extent to which they suffer non-recoverable losses, excluding those losses from the damage measure may reduce the cross-subsidization that could otherwise result (see §5.2.8).

5.3.2. Other measures of monetary damages

The damage measures identified above in section §5.1.2 above have received the most attention in the law and economics literature, but other measures are also sometimes used. As we show, though, most of the economic effects of these alternate measures can be decomposed into one or more of the effects already considered above.

Cost of completion. One recurring issue involves sellers who breach by leaving work unfinished, or by performing work incorrectly, when it would be extremely costly to finish or correct the work. For example, suppose that the buyer would have realized net gains of v from complete performance, while leaving the work in its current state would leave the buyer with net gains of u, and it would now cost $f > v - u$ to finish the work as it should have been performed. If v and u have been measured correctly, this implies that it would be inefficient to finish the work. However, courts vary in their treatment of this case, sometimes allowing the buyer to recover the completion cost f in damages, while sometimes limiting the buyer to recovery of $v - u$.

Awarding f in damages would be inefficient if it led to the work actually being completed. If, however, completion would in fact be inefficient, then the buyer will pocket the damage payment rather than use it to finish the work. The principal effects of awarding f, rather than awarding $v - u$, will, thus, be those discussed previously. The larger remedy will (in general) strengthen sellers' incentives to avoid committing this sort of breach; perhaps strengthening them excessively, if the actual losses caused by the breach are only $v - u$. On the other hand, if $v - u$ actually understates the buyer's losses—say, because some of the benefits from performance are hard to measure, and have therefore been omitted from the court's measure of v—then increasing the remedy to f could improve the seller's incentives in some respects (see Muris, 1983, for a discussion of both effects). Even then, much would depend on the exact nature of the benefits that were excluded, as was also discussed in preceding subsections. For example, if buyers differ in the extent to which they would realize certain benefits, excluding those benefits from the damage measure could reduce any cross-subsidization that might otherwise result (see §5.2.8).

Disgorgement. Similarly, courts occasionally require a breaching party to disgorge any profits he may have earned as a result of the breach, even if those profits exceed the non-breacher's expectation loss, often citing a general principle that no one should profit from his own wrong. This result is not the norm in contract cases, but is reserved for situations in which the breacher is thought to have engaged in bad faith or "willful" breach (an ill-defined notion in the case law) or when the non-breacher is considered to have a property or quasi-property interest in the subject of the contract and is thus entitled to the proceeds of resale, even if he could not have earned such proceeds on his own (see Farnsworth, 1985, for a general discussion).

Disgorgement damages, if assessed with certainty, leave the breaching party indifferent between performance and breach. As such, they entirely eliminate any incentive to breach, which from the viewpoint of efficient breach, is an excessive deterrent. In turn, the absence of efficient breach could also generate excessive reliance by the promisee.

A rationale sometimes articulated in favor of disgorgement damages is that, by removing any incentive for unilateral breach, they encourage a party who would like to escape performance to approach the counterparty and negotiate a modification or release (see, *e.g.*, Friedmann, 1989). Such negotiation may be desirable if the potential breacher would otherwise be uninformed about the size of the counterparty's expectation loss (thus leading him mistakenly to breach), or if the transaction costs of renegotiation are less than the costs that would be occasioned by a lawsuit. The frequent association of disgorgement with the elements of bad faith or willfulness, however, suggests an underlying punitive component to this remedy, which raises the topic of punitive damages generally.

Punitive damages. In general, explicitly punitive damages are rarely imposed in contracts cases. This is in part because breach of contract, though a legal wrong, is typically judged more leniently than would breaches of duty in a tort or property setting, for a number of related reasons. First, because liability is based on a voluntary exchange relationship, the expected costs of breach and of paying damages are likely to be reflected in the contract price, so that efficient breaches work to the *ex ante* advantage of both parties. Second, to the extent that punitive damages are justified by negative externalities imposed on the general community, the concern is less pressing in the ordinary contract. Third, to the extent that punitive damages are justified by a high likelihood of non-detection, this concern is lessened when the parties know each other and are likely to be watchful of proper performance.

Punitive damages are very occasionally imposed, however, in response to breaches of contract that also involve a tort, gross unfairness, or a violation of some public policy. For example, punitive damages have regularly been imposed on insurance companies that refuse without justification to pay valid claims. This result is often defended in terms of the imbalance in the parties' economic power, the particularly difficult circumstances in which such refusal places the insured, and the likelihood that most victims of such opportunism are likely never to seek redress in court—all factors that go to the general case for punitive damages, as laid out in succeeding chapters of this handbook.

5.3.3. Specific performance

In some cases courts order specific performance rather than awarding monetary damages for breach. In effect, this remedy requires the seller (or other breaching party) to perform the contract in full, backed by the threat of fines or even imprisonment if she fails to do so.

If *ex post* renegotiation is impossible, such an order would lead to inefficient performance of the contract in any case where breach was more efficient. As long as the parties can renegotiate, however, they should always be able to avoid this loss by agreeing not to perform, with the gains from non-performance being shared between the parties in accordance with their bargaining strength. But if the buyer has the threat of a remedy of specific performance, thereby requiring the seller to incur the costs of performance, that should allow the buyer to capture more of the gains than he could if his only legal threat were to hold the seller responsible for some smaller monetary remedy.

As a result, when *ex post* renegotiation is possible, the effects of specific performance will be felt in all of the ways discussed above. When renegotiation is costly, specific performance could, in principle, add to those negotiation costs, though it is unclear whether the negotiations required under specific performance will be any more or less costly than those required under any alternative remedy.[91]

In any event, if specific performance results in the buyer being able to negotiate for a larger payment, this will have the same effect as an increase in the expected size of any monetary remedy. For example, the threat of having to make such a payment will strengthen the seller's incentives to take precautions against events that might expose it to such a remedy (Muris, 1982). Moreover, this threat will also increase the buyer's incentives to make reliance investments (Edlin and Reichelstein, 1996). An increase in the expected payment will also alter the risks born by each party, as discussed earlier in §5.2.9. In general, specific performance makes it more likely that the buyer will end up at least as well off as if the contract had been performed (otherwise, he would not consent to any renegotiation), whereas the award of smaller monetary remedies might not leave the buyer this well off. But whether this increase in compensation is desirable, all things considered, depends on all of the effects discussed earlier, as compared to the effects produced by whatever monetary measure of damages the court would award if it did not require specific performance. In this regard, the choice between specific performance and monetary damages has much in common with the choice between injunctive relief and monetary remedies in many other areas of law.[92]

Under common law, specific performance has traditionally been more difficult to obtain than monetary damages, with injunctive relief treated as matter of the court's discretion, and usually being reserved for cases in which damages are deemed insufficient to protect the non-breacher's expectation interest or in other special circumstances (for example, when the goods being traded are unique, when the breacher is insolvent, or when the non-breacher has made relational investments that would be difficult to replace). The most widely cited policy reason for these restrictions is that specific performance is thought to impose greater administrative costs on the legal system, especially in situations where the quality of a coerced performance is costly to verify. But this concern has not prevented civil law systems from making specific performance their remedy of default, even though in many circumstances their courts will still award money damages in substitution for performance (Lando and Rose, 2004).

5.3.4. Remedies expressly stipulated in the contract

As noted earlier (see §5.1.2), the remedies discussed above are usually supplied by the legal system as default remedies to be applied in cases where the contract is silent as

[91] For analyses of this issue see, *e.g.*, Kronman (1978b), Schwartz (1979), Ulen (1984), and Bishop (1985). Most recently, Shavell (2006b) has argued that negotiation costs will be relatively large where the reason for breach is a production cost increase (because in order to negotiate a release, the parties must agree on how to distribute the quasi-rents arising from the cost increase), and resulting), and relatively small where the reason for breach is to sell to an third party (because the buyer can also resell to the third party, and so the quasi-rents are limited to the difference between the buyer's and seller's cost of resale).

[92] There is a much more extensive literature analyzing the analogous issue in property and tort law. See, *e.g.*, Calabresi and Melamed (1972), Kaplow and Shavell (1996), Bebchuk (2001), and Ayres and Goldbart (2003).

to the consequences of breach. In some cases, though, the parties' contract may itself stipulate the remedy that will be required in certain states of the world. Usually this remedy will consist of a monetary payment, but in some cases, the parties may provide for particular actions to be taken, as when a sales contract provides that in the event of a defect in the goods, the seller will be obliged to provide repair or replacement. When such stipulations take a monetary form, they are usually referred to as "stipulated damages" or, at least when they are enforceable, "liquidated damage clauses."[93]

Depending on the amount chosen by the parties to serve as the measure of damages, liquidated damage clauses can produce any and all of the effects described above.[94] For example, clauses that specify a large payment can give sellers a strong incentive to avoid committing a breach, while clauses that specify smaller payments will give buyers less insurance against breaches. The amount of the clause can also affect each side's incentives to rely on the contract, or to take precautions against various risks, as the preceding sections also discussed. Indeed, while the literature summarized in the preceding sections was mostly written to provide guidance to courts or other lawmakers, that literature can just as easily be read as providing guidance to private parties who wish to select an efficient liquidated damage remedy (see Katz, 1996b).

Moreover, in some respects liquidated damage clauses (drafted by the parties) are likely to be superior to general default rules (selected by courts or legislatures). As the preceding sections make clear, most remedies involve trade-offs among various important incentives, and in many cases the contracting parties are better suited than courts to choose the particular trade-off that is best for their own transaction. In markets where parties are heterogeneous (see §5.2.8), liquidated damage clauses can be tailored to particular contracting pairs. More broadly, all of the reasons that support the enforcement of contracts generally (see §2.2) will usually argue for the enforcement of liquidated damage clauses in particular.

It is therefore striking that common-law courts refuse to enforce clauses that set damage amounts that the courts consider excessive (these are typically referred to as "penalty clauses"). This reluctance may be due partly to historical factors, and in particular to the belief (on non-economic grounds) that "penalties" or "enforcement" should be the exclusive province of the legal system, rather than being subject to private control. In addition, though, there may also be economic reasons that—in particular situations—might counsel against enforcement.

For example, in cases where the actual damages from breach turn out to differ significantly from what the parties expected, it is possible—though far from automatic—that a remedy specified by the liquidated damage clause might no longer be welfare max-

[93] Very occasionally, parties will provide for specific performance in their contracts, but courts do not regard themselves as bound by such provisions (although they will take them into account as a factor bearing on their exercise of discretion).

[94] For discussions of these effects in connection with liquidated damage clauses, see Goetz and Scott (1977), Clarkson et al. (1978), Rea (1984), Schwartz (1990), and Edlin and Schwartz (2003).

imizing.[95] In effect, denying enforcement in these cases would be similar to denying enforcement of the contract itself, under the "implied excuse" doctrines of mistake or impracticability, in cases where unforeseen events have significantly altered the contracting parties' situation. It is, however, a matter of debate whether courts have the ability to make such *ex post* adjustments in ways that will in fact improve the parties' *ex ante* welfare. These doctrines are discussed above in §4.4; for discussions focusing specifically on liquidated damage clauses, see Rea (1984) and Schwartz (1990).

In addition, some liquidated damage clauses can affect the welfare of others who are not parties to the contract. In particular, if a seller with market power is concerned with defending its market against entrants, liquidated damage clauses can serve as a commitment to deter competitors from entering (Aghion and Bolton, 1987; Chung, 1991; and Spier and Whinston, 1995). Even in competitive markets, if some or all buyers are unaware of seller's liquidated damage clauses, this imperfect information could create incentives for socially undesirable clauses (just as they could in the case of any contractual clause, see generally §2.3.2).

It is worth noting that there are some situations in which common-law courts are more likely to enforce privately stipulated remedies. In particular, courts are more deferential to liquidated damage clauses that turn out to be undercompensatory *ex post*, as compared to those that turn out to be over-compensatory (perhaps because the latter raise third-party effects where the former do not). They are quite deferential to clauses that disclaim liability for consequential damages and that limit the remedy for breach of warranty to repair or replacement (because such clauses supplement the doctrine of *Hadley v. Baxendale*, of which the courts approve). They are also likelier to enforce liquidated damages that have actually been paid over as an advance deposit, although even in those cases, the breacher may be entitled to restitution of part or all of the deposit to the extent it plainly exceeds the non-breacher's expectation loss.

5.4. *Private enforcement of contracts*

As noted, legal enforcement or its threat are not the only means by which parties are induced to honor their commitments. In this subsection, we briefly consider some of these other means. In §5.4.1, we briefly consider how the costs of using legal enforcement can either distort contracts or cause the parties to dispose of them altogether. In §5.4.2, we take up how repeated interactions or reputational concerns can deter breach. Finally, §5.4.3 discusses various legal doctrines that bear on, and in some cases support, these private alternatives to traditional legal enforcement.

5.4.1. *Enforcement costs*

One reason contracts could fail to be enforced as written is that enforcement requires expenditures by the parties that are either *ex post* incredible or can be anticipated to be so

[95] Observe, however, that if the remedy is a monetary transfer, then welfare is typically assumed to be unaffected by transfers.

large *ex ante* that no contract is written. A partial list of examples is: (1) the agreement is illegal or exceeds the parties' power to contract under the applicable legal system (see §2.5 above); (2) courts cannot verify critical aspects of contractual performance or whether relevant contingencies have arisen (see §4.2 above); (3) litigation is costly in terms of time, risk, or material resources; (4) the defendant may be insolvent or otherwise lack the ability to comply with a judgment; and (5) the dispute arises in the course of an otherwise successful relationship that the parties do not wish to jeopardize. In addition, as a large literature makes clear, two additional reasons are also important in developing countries:[96] (6) the court system operates corruptly; and (7) courts are incapable of enforcing their verdicts, because police are corrupt or unavailable.

A simple model illustrates some of these issues. Suppose that party B employs party A and promises to pay A w upon completion of A's task. Suppose it costs A $k_A > 0$ to have the contract enforced or fight litigation and it costs B $k_B > 0$ for the same. If A cannot recover its enforcement costs from B should she prevail at trial, then B knows he can underpay A by up to k_A without A seeking to enforce the contract (assuming A acts in a coldly rational way).[97] If, however, the parties anticipate this, then they could agree to a nominal wage of $w + k_A$, recognizing that B will underpay by k_A. Somewhat more problematic is non-performance by A. If it is feasible for A to overperform by an amount k_B, then the parties can simply set the performance standard in the contract k_B above what they truly intend. But such overperformance could be infeasible, in which case this trick won't work. Now it could be necessary to add a clause to the contract that A pay B k_B should A be found to have underperformed. To the extent the court refuses to honor that clause, citing the unenforceability of penalty clauses (see §5.3.4), or its inability to verify A's performance adequately, this solution could also fail. If it is impossible to enforce a contract contingent on A's performance, then the parties will either have to forgo contracting or they will have to contract around the problem (*e.g.*, use a revenue-sharing contract to give A better incentives).

More generally, whenever judicial enforcement is likely to occur in less than 100% of the cases, it might in principle be possible to make up for that deficiency by increasing the size of the damage award in those cases that do reach the courts. To take a simple example, if only one out of ten breaches is ever sanctioned by courts, many of the incentive effects could be restored if the damage award that would otherwise be optimal were multiplied by ten in every case.[98] However, courts are usually reluctant to make this adjustment in contract cases, where punitive damages are only rarely awarded. This solution will also be unattractive if either party is risk averse, as it increases the variance

[96] See, for instance, Anderson and Young (2006), Cungu and Swinnen (2003), or, for a more historical perspective, Greif and Kandel (1995).

[97] It is known from research on ultimatum games (*e.g.*, Güth et al., 1982) that players do not always act in a coldly rational way, preferring sometimes to punish others even if the act of punishing is costly to them. See Rabin (1993) for a discussion and analysis.

[98] This effect is often suggested as an economic rationale for punitive damages. See, *e.g.*, Polinsky and Shavell (1998) and Craswell (1999).

of both parties' returns (see §5.2.9). And in more extreme settings, where litigation is infinitely costly (*e.g.*, performance is completely unverifiable or in countries where a reliable court system is unavailable), reliance on court-ordered remedies is bound to fail.

Unfortunately, the issue of credible enforcement is typically ignored in most of the contract design literature.[99] Accordingly, we now turn briefly to enforcement methods that do not requires the participation of courts.

5.4.2. *Self-enforcing contracts*

It has long been understood from the repeated games literature that some agreements are self enforcing in the context of an ongoing relationship.[100] The most prominent example of such "agreements" is tacit collusion among competing firms. That is, recognizing their repeated interaction, firms avoid undercutting each other on price. This "agreement" to keep prices high is enforced by the threat of a price war should any firm undercut.[101]

Within the realm of contracts, there is wide scope for such self-enforcing agreements. Moreover, self-enforcement can substitute for legal enforcement. For instance, in a one-shot game, an employer might choose to renege on a promised wage payment to an employee (recall the discussion in the previous subsection). But in a repeated context, the employee could retaliate by quitting (or possibly engaging in sabotage), which could be sufficiently costly to the employer that he chooses to pay the employee.[102] Of course, the employee could also take the employer to court for nonpayment, but, as we saw above, legal enforcement might not always be credible.

Self enforcement can also complement legal enforcement. For instance, while it might not be credible for a party to enforce a contract legally in a one-shot game, the party might wish to develop a reputation for enforcement in a repeated context—just as in some models of entry deterrence whereby an incumbent firm punishes entrants to establish a reputation for toughness (see, *e.g.*, Tirole, 1988, Chapter 8), a party may want to develop a reputation as someone who can't be cheated (*e.g.*, become known as litigious). Of course, if repetition is what is making legal enforcement of contracts credible, then one should model repetition explicitly in the analysis of the underlying contracting problem if that problem is itself repeated (*i.e.*, the game between an employer and a long-term employee). In other words, an appeal to reputation for enforcement in a static analysis of a contracting problem is most acceptable when the problem is short term (*i.e.*, a given pair meet only once to contract), while the players are long lived (*i.e.*, will play again with others).

[99] Some notable exceptions are Spier (1994) and Krasa and Villamil (2000).

[100] For a review of repeated games see Chapter 2 of Gibbons (1992) or Chapter 5 of Fudenberg and Tirole (1991).

[101] See Chapter 6 of Tirole (1988) for details.

[102] See, also, Thomas and Worrall (1988).

Within the literature, self-enforcing contracts are often known as *relational contracts*. Applications have included quality assurance for experience goods (*i.e.*, goods the quality of which can only be assessed via consumption);[103] incentive schemes;[104] and social contracts within firms.[105] However, explicit models of reputation in the legal literature on contacts are still relatively rare.[106]

In addition, while reputations can, in many markets, provide powerful incentives to perform, in some markets they are less likely to be effective. For example, if performance involves a credence good (the quality of which cannot be observed even after consumption, at least not without expert diagnosis), many breaches may go undetected, with little harm to the breaching party's reputation.[107] In addition, the enforcement of reputations may be privately costly to those who enforce them, thus leading to free-rider problems in enforcement.[108] In other markets, where sellers' histories are not easily discoverable by buyers (or *vice versa*), the incentive effects of reputations may be weakened, though the involvement of kin or ethnic groups or other reputational intermediaries may help in overcoming that difficulty.[109] Finally, for sellers who are on the verge of bankruptcy (or are otherwise reaching their "last period"), the prospect of losing future business may be a very weak constraint at best.

5.4.3. Legal doctrines affecting self-enforcement

As the preceding subsection discussed, reputations are most effective in the context of a repeated game, so that a party who cheats suffers the consequence by losing the benefit of future interactions. In some contractual settings, however, the party who cheats can be made to suffer an extra-legal sanction in connection with the very contract has that been breached, as long as the other party has not yet fully performed his own end of the contract. In such a case, the other party may respond not by filing a lawsuit, but by withholding his own performance. For example, if a seller agrees to deliver goods on credit, but if the buyer discovers (before paying) that the goods are defective, the buyer might respond to this breach by refusing to accept or to pay for the goods.

To be sure, the significance of a threat to suspend performance depends partly on the value that performance has to the other party, but it also depends on how the parties have structured their transaction. To take an extreme case, if the contract calls for the

[103] Klein and Leffler (1981) and Shapiro (1983) are two examples.

[104] Bull (1987), Levin (2003), and MacLeod and Malcomson (1989) are three examples.

[105] See Hermalin (2001) for a survey.

[106] For a qualitative discussion, see Charny (1990). Bernstein (1992, 1996, 2001) has presented several case studies illustrating the operation of reputational enforcement in specialized markets.

[107] See Darby and Karni (1973). Although see Fong (2005) for a more nuanced analysis.

[108] For example, Klein and Leffler (1981) model a market in which buyers, if they even once receive a defective product, follow a flat rule of never purchasing from that seller again. While such a rule does produce desirable incentive effects, it may or may not be rational for individual buyers.

[109] See Landa (1981) and Bernstein (1992) for discussion of kin and ethnic networks. See Mann (1999) on other reputational intermediaries.

seller to make all her deliveries before the buyer pays any of the price, that gives the buyer a good deal of leverage by threatening to withhold payment. On the other hand, if the contract instead calls for the buyer to pay the entire price in advance, before any of the goods have been shipped, this gives the buyer very little leverage, while putting the seller in the happy position of being able to threaten to suspend all of her shipments. As a consequence, parties often negotiate extensively over the exact timing of the various payment and delivery requirements.[110]

In addition, a party's right to suspend his or her own performance may also be regulated by various legal doctrines, as we now discuss.

Rescission. Once one party has committed a breach, the other party may sometimes be able to choose between monetary remedies (typically expectation damages, as discussed earlier in section §5.1.2) and simply walking away from the contract, without collecting any remedy at all. This latter option is usually referred to as termination or rescission. To be sure, if the contract would have been a profitable one for the non-breaching party, that party will usually prefer expectation damages over rescission, for expectation damages should give the non-breacher all of the benefits she would have received from performance (if all of those benefits can be adequately measured, see section §5.2.1 above). But if the contract in question would have been a losing one for the non-breaching party, rescission may be a more valuable remedy, as it allows that party to walk away from what might otherwise be a significant loss.

To complicate matters further, in some cases the non-breaching party may elect to rescind a contract even if she has already performed some part of her own services under the contract. By rescinding the contract, the non-breaching party would give up her right to recover the payment specified in the contract, but she could then sue for restitution to recover a judicially-determined "reasonable value" for her services.[111] As the court's determination of reasonable value need not be limited by the contract price for those services, this remedy could leave the non-breacher with more than she could get under any alternative remedy. Indeed, precisely because rescission is elective for the non-breacher, a rational non-breacher will not choose it unless it is more favorable to her.

However, several legal doctrines limit the use of rescission as a response to the other party's breach. At common law, breach of a service contract allows the non-breaching party to elect rescission only if the breach is "material" or "substantial," a vague test that leaves much to the courts' discretion. By contrast, breach of a contract for the sale of goods is said to allow rescission for any breach whatsoever (the so-called "perfect tender" rule). The Uniform Commercial Code has altered this latter rule, though, by limiting the buyer's right to rescind in cases involving the sale of goods. Under the

[110] See Scott and Triantis (2006). In addition to the sources cited earlier in §5.2.6, brief discussions can also be found in Goetz and Scott (1983), Kull (1991), and Kraus (1994). Note, also, the connection between this and the discussion of option contracts and holdup in §4.3.2.

[111] For other uses of restitution as a remedy, see section §5.1.2 *supra*.

UCC, defective goods in a single shipment of a multi-shipment (or installment) contract do not allow the buyer to rescind the entire contract unless the defect "substantially impairs" the value of the entire contract. Even in contracts calling for only a single installment, the buyer may lose the right to rescind if he fails to reject the installment within a "reasonable" time for inspection. The buyer's right to rescind may also be limited by the seller's right to take a "reasonable" amount of time to "cure" the defect (see the discussion of "cure" below).

While the remedy of rescission has not been analyzed as extensively as other remedies have, many of the effects are similar to those of any other remedy that is more generous to the non-breacher.[112] That is, as noted, the non-breacher will elect rescission only when it is more favorable to her than the other available remedies. Consequently, the availability of rescission should increase the breacher's incentive to perform—just as would any other increase in the size of the likely remedy. Of course, the parties may be able to renegotiate *ex post* to avoid inefficient breach or inefficient performance—but, as with other remedies, the payments that parties must make in *ex post* renegotiations will still affect their *ex ante* investment incentives. Also, if either party is risk averse, the availability of rescission will also affect their risk-bearing costs, again in the same manner as any other increase in the expected size of the remedy.

Cure. As noted in the preceding paragraphs, the buyer's right to rescind a contract for the sale of goods may be limited (under the UCC) by the seller's right to cure any defects in the goods that she delivered. As long as the time specified for delivery of the goods has not yet expired, the seller has complete freedom to try to cure the defects and deliver conforming goods. However, even after the time for delivery has expired, the seller may still have some right to attempt a cure, although this right is subject to various legal limits (many of which are vague). For example, the seller may not take more than a "reasonable" time to effect such a cure; and may only do so when she had "reasonable grounds" to believe that her original, non-conforming delivery would have been acceptable "with or without a monetary allowance" for the defect. In installment contracts, where only the goods in a single shipment were defective, the seller must be able to provide "adequate assurances" that its cure will be successful. And in all cases, it is ultimately up to a court to determine whether the seller's efforts have in fact cured the defect.

In cases where it is clear that the seller can fix the defect, the right to cure serves to limit the effect of the remedy of rescission. For example, if the market price of goods fell significantly after the contract was signed, the buyer might otherwise use a trivial defect—one that could be cured at a *de minimis* cost—to rescind the contract, thus forcing the seller to bear the loss from the market's fall. The economic effects of this use of rescission were discussed above. Clearly, giving the seller a right to cure eliminates those effects.

[112] The earliest economic discussion is Goetz and Scott (1983). Other discussions include Kull (1991) and Kraus (1994).

In other cases, though, it may not be clear (at least initially) whether the seller will ever be successful in curing the defect. For example, many litigated cases involve the sale of cars or houses that seem to be "lemons," whose seller makes repeated but unsuccessful attempts to find and fix the problem.[113] In these cases, a court's interpretation of the right to cure has the effect of determining the point at which the contractual endeavor should be terminated in order to cut the parties' losses. The economic implications of this decision were discussed earlier in §5.2.6.

Conditions and termination clauses. Parties can also use the contract itself to specify (within limits) the conditions under which either or both parties will be released from their obligation to perform. Employment contracts, for example, may allow either the employer or the employee to terminate the relationship at any time, and for any reason. (Indeed, this is the common-law default rule for employment contracts.) Similarly, franchise contracts may specify that the franchise relationship will continue indefinitely unless one or the other party exercises its right to terminate the relationship, often with some advance notice required (*e.g.*, 30 days' notice of termination).

Other contracts may permit one part to terminate the relationship if certain terms of the contract are violated. Technically, contractual clauses whose violation will release one party from part or all of the contract are referred to in law as conditions. By contrast, covenants or promises are clauses whose violation normally leads to some other default remedy. Violation of a covenant will release the other party from the contract only if the breach is found to be "material" (see the preceding discussion of rescission).

To be sure, just as courts sometimes refuse to enforce liquidated damage clauses (see §5.3.4), they also do not always enforce contractual termination provisions. For example, if termination would inflict on the other party a loss that seems to the court to be excessive (a "forfeiture"), courts may refuse to give effect to an express condition, thus prohibiting the other party from terminating the contract. In addition, clauses that purport to give one party the right to terminate a relationship for any reason will sometimes be interpreted more narrowly by courts, who may refuse to permit terminations that are not made "in good faith" (a phrase whose exact content is difficult to pin down). By requiring some degree of judicial approval of a termination decision, these doctrines thus limit the parties' ability to use termination as a self-enforcement technique. Some of the economic effects of these limits are discussed in Klein (1980).

5.5. Other law bearing on contract enforcement

It should also be kept in mind that the enforcement of contracts is often affected by rules and institutions from other fields of law. For example, one common way for parties to

[113] In some cases, the contract may itself specify that the seller has the right to cure, and may even limit the buyer's remedy to accepting the seller's repair or replacement. Even in these cases, though, the UCC allows courts to disregard such a clause (and bring the seller's right to cure to an end) if the seller's inability to cure causes this remedy to "fail its essential purpose."

enhance the likelihood of contractual performance is to offer collateral; and this device is regulated generally by the law of property and specifically by the specialized law of secured transactions (see Schwartz, 1989 and Triantis, 1992). Similarly, contracting parties often enlist third parties as guarantors on their behalf; and the value of such assurance is determined by the law of suretyship (see Katz, 1999).

Conversely, parties' ability to use reputation or repeat dealing as private enforcement devices may be restricted by other fields of law. For example, the law of antitrust generally prohibits concerted boycotts or refusals to deal; and the law of torts may treat some reputation-affecting communications as unfair competition, defamation, or invasion of privacy. To this extent, some private attempts to enforce contracts may be actionable in their own right.

6. Conclusions

This chapter is lengthy and many of our conclusions have already been given. Consequently, we limit ourselves here to a few remarks.

Given the vastness of the literature on the law and economics of contracts, even a survey as long as ours must omit certain topics. One topic that has been omitted is the connection between the literature on contracts and those on torts, takings, and regulation. In particular, much of the economic analysis of regulation takes the view that the regulator and the regulated entity are entering into what is, effectively, an agency contract between the regulator (the principal) and the regulated entity (the agent). See, for instance, Laffont and Tirole (1993). But even torts and takings can be related to contracts if, as some analysis has done, one views the state as seeking to approximate, in some way via law, the contract that it would have liked to have written with the tortfeasor or the owner of the property to be taken if their identity were known in advance.[114]

We have also omitted the entire literature in which the state itself is a party to the contract. Government contracts raises a number of additional issues, including the need for public accountability, risks of corruption and political capture, the problem of establishing credible commitment that will survive changes of governmental regime, and the special difficulties of enforcing contract rights against a sovereign state. For instance, the risk of nations repudiating their debt contracts or abrogating licensing agreements is well documented, especially in the context of developing economies. Creditors or technology providers who contemplate entering into agreements with such governments, accordingly, must find ways to mitigate or insure against such risk. At the same time, citizens and regulators have an interest in preventing state officials from entering into contracts that are not in the public interest *ex ante*. To an extent, these issues relate to our discussion of alternative means of enforcement in §5.4.2 and §5.4.3, but the area is broader than this.

[114] See, *e.g.*, Hermalin (1995) for an application of this approach on the takings issue.

Any chapter of this sort should close with some suggestions for future research. Some suggestions have made already: (i) more economic analysis of non-Anglo-American contract law; (ii) more positive analyses of contract law; (iii) more efforts in modeling to treat monitoring and measurement as endogenous with respect to what information is observable or verifiable; (iv) empirical studies of how courts employ certain rules, such as the parol evidence rule; (v) economic analyses of some of the doctrinal complications associated with the law of contract formation (*e.g.*, promissory estoppel); and (vi) more analysis of the interactions between private and state enforcement of contracts. To this list we would add greater use of new economic paradigms such as behavioral economics. The behavioral paradigm in particular holds out the promise of increased understanding of the phenomenon of bounded rationality, and of legal doctrines that respond to it. Consider, for instance, the literature discussed in §2.3.4 above, as well as more recent work such as that of DellaVigna and Malmendier on issues of contract design and self control and their application to questions of how do health clubs design their contracts (DellaVigna and Malmendier, 2004 and 2006, respectively).

References

Adler, B.E. (1999). “The questionable ascent of Hadley v. Baxendale”. Stanford Law Review 51 (6), 1547–1589.

Aghion, P., Bolton, P. (1987). “Contracts as a barrier to entry”. American Economic Review 77 (3), 388–401.

Aghion, P., Dewatripont, M., Rey, P. (1994). “Renegotiation design with unverifiable information”. Econometrica 62 (2), 257–282.

Aghion, P., Hermalin, B.E. (1990). “Legal restrictions on private contracts can enhance efficiency”. Journal of Law, Economics, and Organization 6 (2), 381–409.

Akerlof, G. (1970). “The market for ‘lemons’: Qualitative uncertainty and the market mechanism”. Quarterly Journal of Economics 84 (3), 488–500.

Anderlini, L., Felli, L. (1994). “Incomplete written contracts: Undescribable states of nature”. Quarterly Journal of Economics 109 (4), 1085–1124.

Anderlini, L., Felli, L., Postlewaite, A. (2003). Should courts always enforce what contracting parties write? Working paper, University of Pennsylvania.

Anderson, J., Young, L. (2006). “Trade and contract enforcement”. Contributions to Economic Analysis and Policy 5 (1), Article 30.

Arrow, K.J. (1963). Social Choice and Individual Values, 2nd edn. Yale University Press, New Haven, CT.

Arrow, K.J., Debreu, G. (1954). “Existence of an equilibrium for a competitive economy”. Econometrica 22 (3), 265–290.

Atiyah, P. (2003). An Introduction to the Law of Contract, 6th edn. Oxford University Press, Oxford.

Ayres, I. (2003). “Valuing modern contract scholarship”. Yale Law Journal 112 (4), 881–902.

Ayres, I., Gertner, R.H. (1989). “Filling gaps in incomplete contracts: An economic theory of default rules”. Yale Law Journal 99 (2), 87–130.

Ayres, I., Gertner, R.H. (1992). “Strategic contractual inefficiency and the optimal choice of legal rules”. Yale Law Journal 101 (4), 729.

Ayres, I., Goldbart, P.M. (2003). “Correlated values in the theory of property and liability rules”. Journal of Legal Studies 32 (1), 121–151.

Ayres, I., Klass, G. (2005). Insincere Promises: The Law of Misrepresented Intent. Yale University Press, New Haven.

Ayres, I., Rasmusen, E. (1993). “Mutual and unilateral mistake in contract law”. Journal of Legal Studies 22 (2), 309–343.

Baird, D., Weisberg, R. (1982). “Rules, standards, and the battle of the forms: A reassessment of 2-207”. Virginia Law Review 68 (6), 1217–1262.

Bar-Gill, O. (2005). “Seduction by plastic”. Northwestern University Law Review 98 (4), 1373–1434.

Bar-Gill, O., Ben-Shahar, O. (2004). “The law of duress and the economics of credible threats”. Journal of Legal Studies 33 (2), 391–430.

Barnett, R.E. (1986). “A consent theory of contract”. Columbia Law Review 86 (2), 269–321.

Barton, J.H. (1972). “The economic basis of damages for breach of contract”. Journal of Legal Studies 1 (2), 277–304.

Bebchuk, L.A. (2001). “Property rights and liability rules: The *ex ante* view of the cathedral”. Michigan Law Review 100 (3), 601–639.

Bebchuk, L.A., Ben-Shahar, O. (2001). “Precontractual reliance”. Journal of Legal Studies 30 (21, pt. 1), 423–457.

Bebchuk, L.A., Shavell, S. (1991). “Information and the scope of liability for breach of contract: The rule of Hadley v. Baxendale”. Journal of Law, Economics, and Organization 7 (2), 284–312.

Ben-Shahar, O. (1999). “The tentative case against flexibility in commercial law”. University of Chicago Law Review 66 (3), 781–820.

Ben-Shahar, O. (2004). “Agreeing to disagree: Filling gaps in deliberately incomplete contracts”. Wisconsin Law Review (2), 389–428.

Ben-Shahar, O. (2005). “An ex ante view of the battle of the forms: Inducing parties to draft reasonable terms”. International Review of Law and Economics 25 (3), 350–370.

Ben-Shahar, O., Bernstein, L. (2000). “The secrecy interest in contract law”. Yale Law Journal 109 (8), 1885–1925.

Bernheim, B.D., Whinston, M.D. (1998). “Incomplete contracts and strategic ambiguity”. American Economic Review 88 (4), 902–932.

Bernstein, L. (1992). “Opting out of the legal system: Extralegal contractual relations in the diamond industry”. Journal of Legal Studies 21 (1), 115–157.

Bernstein, L. (1996). “Merchant law in a merchant court: Rethinking the code’s search for immanent business norms”. University of Pennsylvania Law Review 144 (5), 1765–1821.

Bernstein, L. (1999). “The questionable empirical basis of Article 2’s incorporation strategy: A preliminary study”. University of Chicago Law Review 66 (3), 710–780.

Bernstein, L. (2001). “Private commercial law in the cotton industry: Creating cooperation through rules, norms, and institutions”. Michigan Law Review 99 (7), 1724–1790.

Birmingham, R.L. (1970). “Breach of contract, damage measures, and economic efficiency”. Rutgers Law Review 24 (2), 273–292.

Bishop, W. (1985). “The choice of remedy for breach of contract”. Journal of Legal Studies 14 (2), 299–320.

Blackorby, C., Donaldson, D. (1988). “Cash versus kind, self-selection, and efficient transfers”. American Economic Review 78 (4), 691–700.

Bolton, P., Dewatripont, M. (2005). Contract Theory. MIT Press, Cambridge, MA.

Bonell, M.J. (1997). An International Restatement Of Contract Law: The UNIDROIT Principles Of International Commercial Contracts. Transnational Publishers, Irvington-on-Hudson, NY.

Boolos, G.S., Burgess, J.P., Jeffrey, R.C. (2002). Computability and Logic, 4th edn. Cambridge University Press, Cambridge, England.

Bork, R.H. (1978). The Antitrust Paradox. Basic Books, New York.

Buchanan, J.M. (1970). “In defense of caveat emptor”. University of Chicago Law Review 38 (1), 64–73.

Bull, C. (1987). “The existence of self-enforcing implicit contracts”. Quarterly Journal of Economics 102 (1), 147–160.

Caillaud, B., Hermalin, B.E. (2000a). Hidden action and incentives. Technical report, University of California, Berkeley, http://faculty.haas.berkeley.edu/hermalin/agencyread.pdf.

Caillaud, B., Hermalin, B.E. (2000b). Hidden-information agency. Technical report, University of California, Berkeley, http://faculty.haas.berkeley.edu/hermalin/mechread.pdf.

Calabresi, G., Melamed, A.D. (1972). "Property rules, liability rules, and inalienability: One view of the cathedral". Harvard Law Review 85 (6), 1089–1128.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., Rabin, M. (2003). "Regulation for conservatives: Behavioral economics and the case for asymmetric paternalism". University of Pennsylvania Law Review 151 (3), 1211–1254.

Charny, D. (1990). "Nonlegal sanctions in commercial relationships". Harvard Law Review 104 (2), 373–467.

Charny, D. (1991). "Hypothetical bargains: The normative structure of contract interpretation". Michigan Law Review 89 (7), 1815–1879.

Che, Y.-K., Hausch, D.B. (1999). "Cooperative investments and the value of contracting". American Economic Review 89 (1), 125–147.

Cheung, S.N. (1969). "Transaction costs, risk aversion, and the choice of contractual arrangement". Journal of Law & Economics 12 (1), 23–42.

Chirelstein, M.A. (2006). Concepts and Case Analysis in the Law of Contracts. Foundation Press, New York.

Cho, I.-K., Kreps, D.M. (1987). "Signaling games and stable equilibria". Quarterly Journal of Economics 102 (2), 179–222.

Chung, T.-Y. (1991). "Incomplete contracts, specific investments and risk sharing". Review of Economic Studies 58 (5), 1031–1042.

Clarkson, K.W., Miller, R.L., Muris, T.J. (1978). "Liquidated damages vs. penalties: Sense or nonsense". Wisconsin Law Review (2), 351–390.

Coase, R.H. (1960). "The problem of social cost". Journal of Law and Economics 3 (1), 1–44.

Cook, P.J., Graham, D.A. (1977). "The demand for insurance and protection: The case of irreplaceable commodities". Quarterly Journal of Economics 91 (1), 143–156.

Cooter, R.D. (1982). "The cost of Coase". Journal of Legal Studies 11 (1), 1–31.

Cooter, R.D. (1985). "Unity in tort, contract, and property: The model of precaution". California Law Review 73 (1), 1–51.

Cooter, R.D., Eisenberg, M.A. (1985). "Damages for breach of contract". California Law Review 73 (5), 1432–1481.

Cooter, R.D., Marks, S., Mnookin, R. (1982). "Bargaining in the shadow of the law: A testable model of strategic behavior". Journal of Legal Studies 11 (2), 225–251.

Craswell, R. (1988). "Precontractual investigation as an optimal precaution problem". Journal of Legal Studies 17 (2), 401–436.

Craswell, R. (1989a). "Contract law, default rules, and the philosophy of promising". Michigan Law Review 88 (3), 489–529.

Craswell, R. (1989b). "Performance, reliance, and one-sided information". Journal of Legal Studies 18 (2), 365–401.

Craswell, R. (1990). "Insecurity, repudiation, and cure". Journal of Legal Studies 19 (2), 399–434.

Craswell, R. (1991). "Passing on the costs of legal rules: Efficiency and distribution in buyer-seller relationships". Stanford Law Review 43 (2), 361–398.

Craswell, R. (1993). "Property rules and liability rules in unconscionability and related doctrines". University of Chicago Law Review 60 (1), 1–65.

Craswell, R. (1996a). "Damage multipliers in market relationships". Journal of Legal Studies 25 (2), 463–492.

Craswell, R. (1996b). "Offer, acceptance, and efficient reliance". Stanford Law Review 48 (3), 481–553.

Craswell, R. (1999). "Deterrence and damages: The multiplier principle and its alternatives". Michigan Law Review 97 (7), 2185–2238.

Craswell, R. (2000). "Do trade customs exist?" In: Kraus, J.S., Walt, S.D. (Eds.), The Jurisprudential Foundations of Corporate and Commercial Law. Cambridge University Press, Cambridge.

Craswell, R. (2003). "In that case, what is the question? Economics and the demands of contract theory". Yale Law Journal 112 (4), 903–924.

Craswell, R. (2006). "Taking information seriously: Misrepresentation and nondisclosure in contract law and elsewhere". Virginia Law Review 92 (4), 565–632.

Crocker, K., Masten, S.E. (1988). "Mitigating contractual hazards: Unilateral options and contract length". RAND Journal of Economics 19 (3), 327–343.

Cungu, A., Swinnen, J. (2003). Investment and contract enforcement in transition: Evidence from Hungary. Technical Report 127, LICOS Centre for Transition Economics, Katholieke Universiteit Leuven, Leuven, Belgium.

Dal Bó, E. (in press). "Bribing voters". American Journal of Political Science.

Darby, M.R., Karni, E. (1973). "Free competition and the optimal amount of fraud". Journal of Law and Economics 16 (1), 67–88.

Debreu, G. (1959). Theory of Value. Yale University Press, New Haven, CT.

Dekel, E., Lipman, B.L., Rustichini, A. (1998). "Recent developments in modeling unforeseen contingencies". European Economic Review 42 (3), 523–542.

Dekel, E., Lipman, B.L., Rustichini, A. (2001). "Representing preferences with a unique subjective state space". Econometrica 69 (4), 891–934.

DellaVigna, S., Malmendier, U. (2004). "Contract design and self-control: Theory and evidence". Quarterly Journal of Economics 119 (2), 353–402.

DellaVigna, S., Malmendier, U. (2006). "Paying not to go to the gym". American Economic Review 96 (3), 694–719.

Demski, J.S., Sappington, D.E. (1991). "Resolving double moral hazard problems with buyout agreements". RAND Journal of Economics 22 (2), 232–240.

Diamond, P.A. (1987). "Search theory". In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: A Dictionary of Economics. Macmillan, New York.

Diamond, P.A., Maskin, E. (1979). "An equilibrium analysis of search and breach of contract I: Steady states". Bell Journal of Economics 10 (1), 282–316.

Diamond, P.A., Maskin, E. (1981). "An equilibrium analysis of search and breach of contract II: A non-steady state example". Journal of Economic Theory 25 (2), 165–195.

Dworkin, R.M. (1980). "Is wealth a value?" Journal of Legal Studies 9 (2), 191–226.

Dye, R.A. (1985). "Costly contract contingencies". International Economic Review 26 (1), 233–250.

Easterbrook, F.H. (1984). "The Supreme Court, 1983 term foreword: The court and the economic system". Harvard Law Review 98 (1), 4–60.

Edlin, A.S., Hermalin, B.E. (2000). "Contract renegotiation and options in agency problems". Journal of Law, Economics, and Organization 16 (2), 395–423.

Edlin, A.S., Hermalin, B.E. (2001). "Implementing the first best in an agency relationship with renegotiation: A corrigendum". Econometrica 69 (5), 1391–1395.

Edlin, A.S., Reichelstein, S. (1996). "Holdups, standard breach remedies, and optimal investments". American Economic Review 86 (3), 478–501.

Edlin, A.S., Schwartz, A. (2003). "Optimal penalties in contracts". Chicago-Kent Law Review 78 (1), 33–54.

Eisenberg, M.A. (1995). "The limits of cognition and the limits of contracts". Stanford Law Review 47 (2), 211–259.

Epstein, R. (1975). "Unconscionability: A critical appraisal". Journal of Law & Economics 18, 293–315.

Farnsworth, E.A. (1985). "Your loss or my gain? The dilemma of the disgorgement principle in breach of contract". Yale Law Journal 94 (6), 1339–1393.

Farnsworth, E.A. (2004). Farnsworth on Contracts, 3rd edn. Aspen Publishers, New York.

Farrell, J. (1987a). "Cheap talk, coordination, and entry". RAND Journal of Economics 18 (1), 34–39.

Farrell, J. (1987b). "Information and the Coase theorem". Journal of Economic Perspectives 1 (2), 113–129.

Farrell, J. (1993). "Meaning and credibility in cheap talk games". Games and Economic Behavior 5 (4), 514–531.

Fishman, M.J., Hagerty, K.M. (2003). "Mandatory versus voluntary disclosure in markets with informed and uninformed customers". Journal of Law, Economics, and Organization 19 (1), 45–63.

Fong, Y. (2005). "When do experts cheat and whom do they target?" RAND Journal of Economics 36 (1), 113–130.

Fried, C. (1981). Contract as Promise: A Theory of Contractual Obligation. Harvard University Press, Cambridge, MA.

Friedmann, D. (1989). "The efficient breach fallacy". Journal of Legal Studies 18 (1), 1–24.

Fudenberg, D., Tirole, J. (1990). "Moral hazard and renegotiation in agency contracts". Econometrica 58 (6), 1279–1319.

Fudenberg, D., Tirole, J. (1991). Game Theory. MIT Press, Cambridge, MA.

Fuller, L.L. (1941). "Consideration as form". Columbia Law Review 41 (5), 799–824.

Fuller, L.L., Perdue, W.R. (1936–37). "The reliance interest in contract damages". Yale Law Journal 46 (1), 52–96. January 1937, 46(3), 373–420.

Gergen, M. (1992). "The use of open terms in contract". Columbia Law Review 92 (5), 997–1081.

Gibbons, R. (1992). Game Theory for Applied Economists. Princeton University Press, Princeton, NJ.

Goetz, C.J., Scott, R.E. (1977). "Liquidated damages, penalties and the just compensation principle: Some notes on an enforcement model and a theory of efficient breach". Columbia Law Review 77 (4), 554–594.

Goetz, C.J., Scott, R.E. (1980). "Enforcing promises: An examination of the basis of contract". Yale Law Journal 89 (7), 1261–1322.

Goetz, C.J., Scott, R.E. (1981). "Principles of relational contracts". Virginia Law Review 67 (6), 1089–1150.

Goetz, C.J., Scott, R.E. (1983). "The mitigation principle: Toward a general theory of contractual obligation". Virginia Law Review 69 (6), 967–1024.

Goetz, C.J., Scott, R.E. (1985). "The limits of expanded choice: An analysis of the interactions between express and implied contract terms". California Law Review 73 (2), 261–322.

Goldberg, V.P. (1998). "Bloomer Girl revisited or how to frame an unmade picture". Wisconsin Law Review 1998 (4), 1051–1084.

Goldberg, V.P. (2000). "In search of best efforts: Reinterpreting Bloor v. Falstaff". St. Louis University Law Journal 44 (4), 1465–1485.

Goldberg, V.P., Erickson, J.R. (1987). "Quantity and price adjustment in long-term contracts: A case study of petroleum coke". Journal of Law & Economics 30 (2), 369–398.

Greif, A., Kandel, E. (1995). "Contract enforcement institutions: Historical perspective and current status in Russia". In: Lazear, E.P. (Ed.), Economic Transition in Eastern Europe and Russia Realities of Reform. Hoover Institution Press, Stanford, CA, pp. 291–321.

Grossman, S.J. (1981). "The informational role of warranties and private disclosure about product quality". Journal of Law and Economics 21, 461–483.

Grossman, S.J., Hart, O.D. (1983). "An analysis of the principal-agent problem". Econometrica 51 (1), 7–46.

Grossman, S.J., Hart, O.D. (1986). "The costs and benefits of ownership: A theory of vertical and lateral integration". Journal of Political Economy 94 (4), 691–719.

Güth, W., Schmittberger, R., Schwarze, B. (1982). "An experimental analysis of ultimatum bargaining". Journal of Economic Behavior and Organization 3 (4), 367–388.

Hadfield, G.K. (1990). "Problematic relations: Franchising and the law of incomplete contracts". Stanford Law Review 42 (4), 927–992.

Hadfield, G.K. (1994). "Judicial competence and the interpretation of incomplete contracts". Journal of Legal Studies 23 (1), 159–184.

Harris, M., Raviv, A. (1978). "Some results on incentive contracts with applications to education and employment, health insurance, and law enforcement". American Economic Review 68 (1), 20–30.

Hart, O.D. (1987). "Incomplete contracts". In: Eatwell, J., Milgate, M., Newman, P. (Eds.), Allocation, Information, and Markets. W.W. Norton & Co., New York.

Hart, O.D. (1989). "An economist's perspective on the theory of the firm". Columbia Law Review 89 (7), 1757–1774.

Hart, O.D., Moore, J.H. (1998). "Default and renegotiation: A dynamic model of debt". Quarterly Journal of Economics 113 (1), 1–41.

Hart, O.D., Moore, J.H. (1999). "Foundations of incomplete contracts". Review of Economic Studies 66 (1), 115–138.

Hart, O.D., Moore, J.H. (2004). Agreeing now to agree later: Contracts that rule out but do not rule in. Harvard Law and Economics Discussion Paper No 465.

Hermalin, B.E. (1995). "An economic analysis of takings". Journal of Law, Economics, and Organization 11 (1), 64–86.

Hermalin, B.E. (2001). “Economics and corporate culture”. In: Cooper, C.L., Cartwright, S., Earley, P.C. (Eds.), The International Handbook of Organizational Culture and Climate. Wiley, Chichester, England.

Hermalin, B.E. (2002). “Adverse selection, short-term contracting, and the underprovision of on-the-job training”. Contributions to Economic Analysis & Policy 1 (1). Article 5.

Hermalin, B.E., Katz, M.L. (1991). “Moral hazard and verifiability: The effects of renegotiation in agency”. Econometrica 59 (6), 1735–1753.

Hermalin, B.E., Katz, M.L. (1993). “Judicial modification of contracts between sophisticated parties: A more complete view of incomplete contracts and their breach”. Journal of Law, Economics, and Organization 9 (2), 230–255.

Hermalin, B.E., Katz, M.L. (2006). “Privacy, property rights, & efficiency: The economics of privacy as secrecy”. Quantitative Marketing and Economics 4 (3), 209–239.

Hillman, R.A. (1979). “Policing contract modifications under the UCC: Good faith and the doctrine of economic duress”. Iowa Law Review 64 (4), 849–902.

Holmes, O.W. (1897). “The path of the law”. Harvard Law Review 10 (8), 457–478.

Holmstrom, B. (1982). “Moral hazard in teams”. Bell Journal of Economics 13 (2), 324–340.

Honnold, J.O. (1999). Uniform Law For International Sales Under The 1980 United Nations Convention. Kluwer Law International, Cambridge, MA.

Jackson, T.H. (1978). “Anticipatory repudiation and the temporal element of contract law: An economic inquiry into contract damages in cases of prospective nonperformance”. Stanford Law Review 31 (1), 69–119.

Johnston, J.S. (1990). “Strategic bargaining and the economic theory of contract default rules”. Yale Law Journal 100 (3), 615–664.

Jolls, C.M. (1997). “Contracts as bilateral commitments: A new perspective on contract modification”. Journal of Legal Studies 26 (1), 203–237.

Jolls, C.M., Sunstein, C.R., Thaler, R. (1998). “A behavioral approach to law and economics”. Stanford Law Review 50 (5), 1471–1550.

Joskow, P.L. (1987). “Contract duration and relationship-specific investments: Empirical evidence from coal markets”. American Economic Review 77 (1), 168–185.

Kahan, M., Klausner, M. (1997). “Standardization and innovation in corporate contracting (or the economics of boilerplate)”. Virginia Law Review 83 (4), 713–770.

Kaplow, L. (1992). “Rules versus standards: An economic analysis”. Duke Law Journal 42 (3), 557–629.

Kaplow, L. (1995). “A model of the optimal complexity of legal rules”. Journal of Law, Economics, & Organization 11 (1), 150–163.

Kaplow, L., Shavell, S. (1996). “Property rules versus liability rules: An economic analysis”. Harvard Law Review 109 (4), 713–790.

Katz, A.W. (1990a). “The strategic structure of offer and acceptance: Game theory and the law of contract formation”. Michigan Law Review 89 (2), 215–295.

Katz, A.W. (1990b). “Your terms or mine? The duty to read the fine print in contracts”. RAND Journal of Economics 21 (4), 518–537.

Katz, A.W. (1993). “Transaction costs and the legal mechanics of contract formation: When should silence in the face of an offer be construed as acceptance?” Journal of Law, Economics, and Organization 9 (1), 77–97.

Katz, A.W. (1996a). “Positivism and the separation of law and economics”. Michigan Law Review 94 (7), 2229–2269.

Katz, A.W. (1996b). “Taking private ordering seriously”. University of Pennsylvania Law Review 144 (5), 1745–1763.

Katz, A.W. (1996c). “When should an offer stick? The economics of promissory estoppel in preliminary negotiations”. Yale Law Journal 105 (5), 1249–1309.

Katz, A.W. (1999). “An economic analysis of the guaranty contract”. University of Chicago Law Review 66 (1), 47–116.

Katz, A.W. (2004). “The economics of form and substance in contract interpretation”. Columbia Law Review 103 (2), 496–538.

Kelman, S. (1981). What Price Incentives? Economists and the Environment. Auburn House Publishing Co., Boston.

Kennedy, D. (1982). "Distributive and paternalist motives in contract and tort law, with special reference to compulsory terms and unequal bargaining power". Maryland Law Review 41 (4), 563–658.

Kenney, R.W., Klein, B. (1983). "The economics of block booking". Journal of Law & Economics 26 (3), 497–540.

Kenney, R.W., Klein, B. (2000). "How block booking facilitated self-enforcing film contracts". Journal of Law & Economics 43 (2), 427–435.

Kessler, F. (1943). "Contracts of adhesion—some thoughts about freedom of contract". Columbia Law Review 43 (5), 629–642.

Klein, B. (1980). "Transaction cost determinants of unfair contractual arrangements". American Economic Review 70 (2), 356–360.

Klein, B., Leffler, K. (1981). "The role of market forces in assuring contractual performance". Journal of Political Economy 89 (4), 615–641.

Klein, B., Murphy, K. (1988). "Vertical restraints as contract enforcement mechanisms". Journal of Law & Economics 31 (2), 265–297.

Kornhauser, L.A. (1983). "Reliance, reputation, and breach of contract". Journal of Law and Economics 26 (2), 691–706.

Korobkin, R.B. (1998a). "Inertia and preference in contract negotiation: The psychological power of default rules and form terms". Vanderbilt Law Review 51 (6), 1583–1652.

Korobkin, R.B. (1998b). "The status quo bias and contract default rules". Cornell Law Review 83 (3), 608–687.

Korobkin, R.B., Ulen, T.S. (2000). "Law and behavioral science: Removing the rationality assumption from law and economics". California Law Review 88 (4), 1051–1142.

Krasa, S., Villamil, A.P. (2000). "Optimal contracts when enforcement is a decision variable". Econometrica 68 (1), 119–134.

Kraus, J.S. (1994). "Decoupling sales law from the acceptance-rejection fulcrum". Yale Law Journal 104 (1), 129–205.

Kreps, D.M. (1990). A Course in Microeconomic Theory. Princeton University Press, Princeton, NJ.

Kronman, A.T. (1978a). "Mistake, disclosure, information, and the law of contracts". Journal of Legal Studies 7 (1), 1–34.

Kronman, A.T. (1978b). "Specific performance". University of Chicago Law Review 45 (2), 351–382.

Kronman, A.T. (1987). "Living in the law". University of Chicago Law Review 54 (3), 835–876.

Kull, A. (1991). "Mistake, frustration, and the windfall principle of contract remedies". Hastings Law Journal 43 (1), 1–55.

Laffont, J.-J. (1988). Fundamentals of Public Economics. MIT Press, Cambridge, MA.

Laffont, J.-J., Martimont, D. (2002). The Theory of Incentives. Princeton University Press, Princeton, NJ.

Laffont, J.-J., Tirole, J. (1988). "The dynamics of incentive contracts". Econometrica 56 (5), 1153–1175.

Laffont, J.-J., Tirole, J. (1993). A Theory of Incentives in Procurement and Regulation. MIT Press, Cambridge, MA.

Landa, J.T. (1981). "A theory of the ethnically homogeneous middleman group: An institutional alternative to contract law". Journal of Legal Studies 10 (2), 349–362.

Lando, H., Rose, C. (2004). "On the enforcement of specific performance in civil law countries". International Review of Law and Economics 24 (4), 473–487.

Lando, O., Beale, H. (Eds.) (2002). Contract Law. Kluwer Law International, The Hague.

Levin, J. (2003). "Relational incentive contracts". American Economic Review 93 (3), 835–857.

Llewellyn, K.N. (1960). The Common Law Tradition: Deciding Appeals. Little, Brown and Co., Boston.

MacLeod, W.B., Malcolmson, J.M. (1993). "Investments, holdup, and the form of market contracts". American Economic Review 83 (4), 811–837.

MacLeod, W.B., Malcomson, J. (1989). "Implicit contracts, incentive compatibility, and involuntary unemployment". Econometrica 56 (2), 447–480.

Macneil, I.R. (1982). "Efficient breach of contract: Circles in the sky". Virginia Law Review 68 (5), 947–969.

Mankiw, N.G., Whinston, M.D. (1986). "Free entry and social inefficiency". RAND Journal of Economics 17 (1), 48–58.

Mann, R.J. (1997a). "Explaining the pattern of secured credit". Harvard Law Review 110 (3), 625–683.

Mann, R.J. (1997b). "The role of secured credit in small-business lending". Georgetown Law Journal 86 (1), 1–44.

Mann, R.J. (1999). "Verification institutions in financing transactions". Georgetown Law Journal 87 (7), 2225–2272.

Mas-Colell, A., Whinston, M.D., Green, J.R. (1995). Microeconomic Theory. Oxford University Press, Oxford.

Maskin, E., Tirole, J. (1999). "Unforeseen contingencies and incomplete contracts". Review of Economic Studies 66 (1), 83–114.

Masten, S.E., Snyder, E.A. (1993). "United States versus United Shoe Machinery Corporation: On the merits". Journal of Law & Economics 36 (1), 33–70.

Matthews, S., Moore, J.H. (1987). "Monopoly provision of quality and warranties: An exploration in the theory of multidimensional screening". Econometrica 55 (2), 441–467.

Matthews, S., Postlethwaite, A. (1985). "Quality testing and disclosure". RAND Journal of Economics 16 (3), 328–340.

Medema, S.G., Zerbe, R.O. (2000). "The Coase theorem". In: Bouckaert, B., De Geest, G. (Eds.), Encyclopedia of Law and Economics, vol. 1. Edward Elgar, Cheltenham, England, pp. 836–892.

Menell, P.S. (1983). "A note on private versus social incentives to sue in a costly legal system". Journal of Legal Studies 12 (1), 41–52.

Muris, T.J. (1982). "The costs of freely granting specific performance". Duke Law Journal (6), 1053–1069.

Muris, T.J. (1983). "Cost of completion or diminution in market value: The relevance of subjective value". Journal of Legal Studies 12 (2), 379–400.

Myerson, R.B. (2004). "Justice, institutions, and multiple equilibria". Chicago Journal of International Law 5 (1), 91–108.

Nash, J.F. (1950). "The bargaining problem". Econometrica 18 (2), 155–162.

Nash, J.F. (1951). "Noncooperative games". Annals of Mathematics 54 (2), 286–295.

Nöldeke, G., Schmidt, K.M. (1995). "Option contracts and renegotiation: A solution to the hold-up problem". RAND Journal of Economics 26 (2), 163–179.

Nöldeke, G., Schmidt, K.M. (1998). "Sequential investments and options to own". RAND Journal of Economics 29 (4), 633–653.

Okun, A.M. (1975). Equality and Efficiency: The Big Tradeoff. The Brookings Institution, Washington, DC.

Pildes, R., Anderson, E. (1990). "Slinging arrows at democracy: Social choice theory, value pluralism, and democratic politics". Columbia Law Review 90 (8), 2121–2214.

Pirrong, S.C. (1993). "Contracting practices in bulk shipping markets: A transactions cost explanation". Journal of Law & Economics 36 (2), 937–976.

Polinsky, A.M. (1974). "Economic analysis as a potentially defective product: A buyer's guide to Posner's Economic Analysis of Law". Harvard Law Review 87 (8), 1655–1681.

Polinsky, A.M. (1983). "Risk sharing through breach of contract remedies". Journal of Legal Studies 12 (2), 427–444.

Polinsky, A.M., Shavell, S. (1998). "Punitive damages: An economic analysis". Harvard Law Review 111 (4), 869–962.

Posner, E.A. (1995). "Contract law in the welfare state: A defense of the unconscionability doctrine, usury laws, and related limitations on the freedom to contract". Journal of Legal Studies 24 (2), 283–319.

Posner, E.A. (1998). "The parol evidence rule, the plain meaning rule, and the principles of contractual interpretation". University of Pennsylvania Law Review 146 (2), 533–577.

Posner, E.A. (2003a). "Economic analysis of contract law after three decades: Success or failure?" Yale Law Journal 112 (4), 829–880.

Posner, E.A. (2006). "There are no penalty default rules in contract law". Florida State University Law Review 33 (3), 563–588.

Posner, R.A. (1971). "Taxation by regulation". Bell Journal of Economics 2 (1), 22–50.

Posner, R.A. (1976). Antitrust Law: An Economic Perspective. University of Chicago Press, Chicago.

Posner, R.A. (1977). "Gratuitous promises in economics and law". Journal of Legal Studies 6 (2), 411–426.

Posner, R.A. (2003). Economic Analysis of Law, 6th edn. Little, Brown & Co., Boston.

Posner, R.A., Rosenfield, A.M. (1977). "Impossibility and related doctrines in contract law: An economic analysis". Journal of Legal Studies 6 (1), 83–118.

Priest, G.L. (1978). "Breach and remedy for the tender of nonconforming goods under the Uniform Commercial Code: An economic approach". Harvard Law Review 91 (5), 960–1001.

Quillen, G.D. (1988). "Contract damages and cross-subsidization". Southern California Law Review 61 (4), 1125–1141.

Rabin, M. (1993). "Incorporating fairness into game theory and economics". American Economic Review 83 (5), 1281–1302.

Rabin, M. (1998). "Psychology and economics". Journal of Economic Literature 36 (1), 11–46.

Rachlinski, J.J. (1998). "A positive psychological theory of judging in hindsight". University of Chicago Law Review 65 (2), 571–625.

Radin, M.J. (1987). "Market-inalienability". Harvard Law Review 100 (8), 1849–1937.

Rakoff, T. (1983). "Contracts of adhesion: An essay in reconstruction". Harvard Law Review 96 (6), 1173–1284.

Rasmusen, E.B. (2001). "Explaining incomplete contracts as the result of contract-reading costs". Advances in Economic Analysis & Policy 1 (1), Article 2.

Rasmusen, E.B., Ramseyer, J.M., Wiley, J.S. (1991). "Naked exclusion". American Economic Review 81 (5), 1137–1145.

Rea, S.A. (1982). "Nonpecuniary loss and breach of contract". Journal of Legal Studies 11 (1), 35–53.

Rea, S.A. (1984). "Efficiency implications of penalties and liquidated damages". Journal of Legal Studies 13 (1), 147–167.

Rogerson, W.P. (1984). "Efficient reliance and damage measures for breach of contract". RAND Journal of Economics 15 (1), 39–53.

Rogerson, W.P. (1992). "Contractual solutions to the hold-up problem". Review of Economic Studies 59 (4), 777–793.

Rubinstein, A. (1998). Modeling Bounded Rationality. Zeuthen Lecture Book series. MIT Press, Cambridge, MA.

Samuelson, W. (1985). "A comment on the Coase theorem". In: Roth, A.E. (Ed.), Game-Theoretic Models of Bargaining. Cambridge University Press, Cambridge, pp. 321–339.

Schelling, T.C. (1960). The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Schlechtreim, P. (1998). Commentary on the UN Convention on the International Sale of Goods. Clarendon Press, Oxford.

Schwartz, A. (1979). "The case for specific performance". Yale Law Journal 89 (2), 271–306.

Schwartz, A. (1989). "A theory of loan priorities". Journal of Legal Studies 18 (2), 209–261.

Schwartz, A. (1990). "The myth that promisees prefer supracompensatory remedies: An analysis of contracting for damage measures". Yale Law Journal 100 (2), 369–407.

Schwartz, A. (1992). "Relational contracts in the courts: An analysis of incomplete agreements and judicial strategies". Journal of Legal Studies 21 (2), 271–318.

Schwartz, A., Scott, R.E. (2004). "Contract theory and the limits of contract law". Yale Law Journal 113 (3), 541–620.

Schwartz, A., Watson, J. (2003). "The law and economics of costly contracting". Journal of Law, Economics, and Organization 20 (1), 2–31.

Schwartz, A., Wilde, L.L. (1979). "Intervening in markets on the basis of imperfect information: A legal and economic analysis". University of Pennsylvania Law Review 127 (3), 630–682.

Schweizer, U. (1988). "Externalities and the Coase theorem: Hypothesis or result?" Journal of Institutional and Theoretical Economics 144 (2), 245–266.

Scott, R.E. (1986). "A relational theory of secured financing". Columbia Law Review 86 (5), 901–977.

Scott, R.E. (2000). “The case for formalism in relational contract”. Northwestern University Law Review 94 (3), 847–876.

Scott, R.E. (2002). “The rise and fall of Article 2”. Louisiana Law Review 62 (4), 1009–1064.

Scott, R.E., Triantis, G. (2006). “Anticipating litigation in contract design”. Yale Law Journal 115 (4), 814–879.

Segal, I.R. (1999a). “Complexity and renegotiation: A foundation for incomplete contracts”. Review of Economic Studies 66 (1), 57–82.

Segal, I.R. (1999b). “Contracting with externalities”. Quarterly Journal of Economics 114 (2), 337–388.

Segal, I.R., Whinston, M.D. (2000). “Naked exclusion: Comment”. American Economic Review 90 (1), 296–309.

Segal, I.R., Whinston, M.D. (2003). “Robust predictions for bilateral contracting with externalities”. Econometrica 71 (3), 757–791.

Sen, A. (1970). “The impossibility of a Paretian liberal”. Journal of Political Economy 78 (1), 152–157.

Shapiro, C. (1983). “Premiums for high quality products as returns to reputation”. Quarterly Journal of Economics 98 (4), 659–680.

Shavell, S. (1980). “Damage measures for breach of contract”. Bell Journal of Economics 11 (2), 466–490.

Shavell, S. (1984). “The design of contracts and remedies for breach”. Quarterly Journal of Economics 99 (1), 121–148.

Shavell, S. (1991). “An economic analysis of altruism and deferred gifts”. Journal of Legal Studies 20 (2), 401–421.

Shavell, S. (1994). “Acquisition and disclosure of information prior to sale”. RAND Journal of Economics 25 (1), 20–36.

Shavell, S. (2006a). “On the writing and the interpretation of contracts”. Journal of Law, Economics, and Organization 22, 289–314.

Shavell, S. (2006b). “Specific performance versus damages for breach of contract: An economic analysis”. Texas Law Review 84 (4), 831–876.

Shavell, S. (in press). “Contracts, holdup, and legal intervention”. Journal of Legal Studies.

Shavell, S., Kaplow, L. (2002). Fairness versus Welfare. Harvard University Press, Cambridge, MA.

Shepard, A. (1987). “Licensing to enhance demand for new technologies”. RAND Journal of Economics 18 (3), 360–368.

Simon, H.A. (1972). “Theories of bounded rationality”. In: McGuire, C., Radner, R. (Eds.), Decision and Organization: A volume in Honor of Jacob Marschak. North Holland, Amsterdam.

Spence, A.M. (1973). “Job market signaling”. Quarterly Journal of Economics 87 (3), 355–374.

Spence, A.M. (1975). “Monopoly, quality, and regulation”. Bell Journal of Economics 6 (2), 417–429.

Spier, K.E. (1992). “Incomplete contracts and signalling”. RAND Journal of Economics 23 (3), 432–443.

Spier, K.E. (1994). “Settlement bargaining and the design of damage awards”. Journal of Law, Economics, and Organization 10 (1), 84–95.

Spier, K.E., Whinston, M.D. (1995). “On the efficiency of privately stipulated damages for breach of contract: Entry barriers, reliance, and renegotiation”. RAND Journal of Economics 26 (2), 180–202.

Sutton, J. (1986). “Non-cooperative bargaining theory: An introduction”. Review of Economic Studies 53 (5), 709–724.

Thomas, J., Worrall, T. (1988). “Self-enforcing wage contracts”. Review of Economic Studies 55 (4), 541–553.

Tirole, J. (1988). The Theory of Industrial Organization. MIT Press, Cambridge, MA.

Tirole, J. (1999). “Incomplete contracts: Where do we stand?” Econometrica 674, 741–781.

Tobin, J. (1970). “On limiting the domain of inequality”. Journal of Law & Economics 13 (2), 263–277.

Townsend, R.M. (1979). “Optimal contracts and competitive markets with costly state verification”. Journal of Economic Theory 20, 265–293.

Treitel, G. (Ed.) (2003). The Law of Contract, 11th edn. Sweet & Maxwell, London.

Triantis, A.J., Triantis, G.G. (1998). “Timing problems in contract breach decisions”. Journal of Law and Economics 41 (1), 163–207.

Triantis, G.G. (1992). “Secured debt under conditions of imperfect information”. Journal of Legal Studies 21 (1), 225–258.

Tullock, G. (2005). The Rent-Seeking Society. Liberty Fund, Indianapolis.

Ulen, T.S. (1984). “The efficiency of specific performance: Toward a unified theory of contract remedies”. Michigan Law Review 83 (2), 341–403.

Varian, H.R. (1992). Microeconomic Analysis, 3rd edn. W.W. Norton, New York.

Walras, L. (1874). Eléments d’économie politique pure. L. Corbaz, Lausanne.

Webb, D.C. (1992). “Two-period financial contracts with private information and costly state verification”. Quarterly Journal of Economics 107 (3), 1113–1123.

Weinstein, M. (1998). “Profit sharing contracts in Hollywood: Evolution and analysis”. Journal of Legal Studies 27 (1), 67–112.

White, J.J., Summers, R.S. (2000). Uniform Commercial Code, 5th edn. West Group, St. Paul.

White, M.J. (1988). “Contract breach and contract discharge due to impossibility: A unified theory”. Journal of Legal Studies 17 (2), 353–376.

Williamson, O.E. (1975). Markets and Hierarchies, Analysis and Antitrust Implications: A Study in the Economics of Internal Organization. The Free Press, New York.

Willig, R.D. (1976). “Consumer surplus without apology”. American Economic Review 66 (4), 589–597.

Wittgenstein, L. (1922). Tractatus Logico-Philosophicus. Routledge and Kegan Paul, London.

Wittman, D. (1981). “Optimal pricing of sequential inputs: Last clear chance, mitigation of damages, and related doctrines in the law”. Journal of Legal Studies 10 (1), 65–91.

Wolcher, L.E. (1989). “Price discrimination and inefficient risk allocation under the rule of *Hadley v. Baxendale*”. Research in Law and Economics 9, 9–31.

Chapter 2

LIABILITY FOR ACCIDENTS

STEVEN SHAVELL[*]

School of Law, Harvard University, and National Bureau of Economic Research

Contents

[*] Research support from the John M. Olin Center for Law, Economics, and Business at Harvard Law School is gratefully acknowledged.

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell
© 2007 Elsevier B.V. All rights reserved
DOI: 10.1016/S1574-0730(07)01002-X

Abstract

This is a survey of legal liability for accidents. Three general aspects of accident liability are addressed. The first is the effect of liability on incentives, both whether to engage in activities (for instance, whether to drive) and how much care to exercise (at what speed to travel) to reduce risk when so doing. The second general aspect concerns risk-bearing and insurance, for the liability system acts as an implicit insurer for accident victims and it imposes risk on potential injurers (because they may have to pay judgments to victims). In this regard, victims' accident insurance and injurers' liability insurance are taken into account. The third general aspect of accident liability is its administrative expense, comprising the cost of legal services, the value of litigants' time, and the operating cost of the courts. A range of subtopics is considered, including product liability, causation, punitive damages, the judgment-proof problem, vicarious liability, and nonpecuniary harm. Liability is also compared to other methods of controlling harmful activities, notably, to corrective taxation and to regulation.

Keywords

law and economics, liability, accident(s), tort(s), insurance, product liability, externalities, information, regulation, litigation, suit, trial, court(s)

JEL classification: D00, D6, D8, H1, H8, K00, K13, L5

1. Introduction

The subject of this chapter is liability for accidents, by which is meant the law determining when the victim of an accident is entitled to recover losses from the injurer. This body of law, included in what is known as tort law, governs, for example, when the victim of an automobile accident can collect from the driver who harmed him, when the victim of pollution can secure compensation from the polluter, or when the victim of an adverse reaction to a drug can obtain a judgment from its manufacturer.

Three aspects of accident liability will be addressed. The first is its effect on incentives, both whether to engage in activities (for instance, whether to drive) and how much care to exercise (at what speed to travel) to reduce risk when so doing. The second aspect concerns risk-bearing and insurance, for the liability system acts as an implicit insurer for victims and it imposes risk on potential injurers (because they may have to pay victims). In this regard, victims' accident insurance and injurers' liability insurance will be taken into account. The third aspect of accident liability is its administrative expense, comprising the cost of legal services, the value of litigants' time, and the operating cost of the courts.[1]

The chapter will begin in Part A with the central theory of accident liability, where the main points about incentives, risk-bearing, and administrative costs will be presented. Then in Part B a variety of subsidiary topics and issues will be discussed, including liability of firms, the judgment-proof problem (inability to pay fully for harm done), causation, nonmonetary losses, and vicarious liability. The sections in Part B can be read more or less independently of one another; they require only an understanding of Part A. Last, Part C will compare liability to other methods of controlling harmful activities, such as corrective (Pigouvian) taxes and regulation.

Economic analysis of accident liability began with mainly informal contributions of a number of legal scholars, notably, Calabresi (1970) and Posner (1972), and has been developed since then, in large part by economists, using the standard methods of microeconomics.

Before proceeding, it may be remarked that in view of the importance of liability in reality (its virtual omnipresence) and its appeal in theory relative to other means of controlling harmful externalities (addressed in Part C), one has the sense that it deserves to receive greater attention from economists than it has to this point.

Part A: Central theory of liability

In this part on the principal theory of liability, the two basic rules of liability will be considered, strict liability and the negligence rule. Under strict liability, an injurer must always pay for harm due to an accident that he causes. Under the negligence rule, an

[1] This chapter is mainly theoretical; for empirical work on liability, see Kessler and Rubinfeld (2007).

injurer must pay for harm caused only when he is found negligent, that is, only when his level of care was less than a standard of care chosen by the courts, often referred to as "due care." (There are various versions of these rules that depend on whether victims' care was insufficient, as will be discussed below.) In fact, the negligence rule is the dominant form of liability; strict liability is reserved mainly for certain especially dangerous activities (such as the use of explosives).[2] The amount paid by a liable party under a liability rule is often referred to as "damages," and unless otherwise noted, damages will be assumed to equal the harm caused in an accident.

2. Incentives

In order to focus on liability and incentives, it is assumed in this section that victims and injurers are risk neutral. Further, it is assumed that they are strangers to one another, or at least are not in a contractual relationship. Additionally, it is assumed for simplicity that victims and injurers are individuals as opposed to firms (although most of what is said carries over to the context of firms; see section 5).

To begin with, accidents are assumed to be unilateral in nature: only injurers can influence risks. Then bilateral accidents are considered; in these accidents victims as well as injurers affect risks. As noted above, two types of decisions of parties are examined: concerning their level of care (or precautions) when engaging in an activity; and concerning their level of activity. First, the choice of care alone will be studied; then both care and activity level will be investigated.

2.1. Unilateral accidents and levels of care

Here the assumption is that injurers alone can reduce risk by choosing a level of care. Let x be expenditures on care (or the value of effort devoted to it) and $p(x)$ be the probability of an accident that causes harm h, where p is declining and convex in x. Assume that the social objective is to minimize total expected costs, $x + p(x)h$, and let x^* denote the optimal x.

Under strict liability, injurers are required to pay damages equal to h whenever an accident occurs, and they bear the cost of care x. Thus, they minimize $x + p(x)h$; accordingly, they choose x^*.

Under the negligence rule, suppose that the due care level, denoted $\underline{x}$, is set equal to x^*. Hence, under the rule an injurer who causes harm will be found negligent and have to pay h if $x < x^*$ but will not be found negligent and will not have to pay anything if $x \geq x^*$. It follows that the injurer will choose x^*: Clearly, he will not choose x greater than x^*, for that will cost him more and he will escape liability for negligence by choosing merely x^*. Moreover, he will not choose $x < x^*$, for then he would be

[2] Dobbs (2000).

liable, implying that his expenses would exceed $x^* + p(x^*)h$, which is greater than or equal to x^*.

Thus, under both forms of liability, strict liability and the negligence rule, injurers are induced to take optimal care. These fundamental results were first shown by Brown (1973). Several comments may be made about them.

(a) The informational requirements imposed on the courts are different under the two rules. Under strict liability, the courts need only to observe the harm h. In contrast, under the negligence rule, the courts need to know more: they must calculate optimal care x^*, observe actual care x, and also observe harm h. (However, the informational advantage of strict liability is attenuated when strict liability with a defense of contributory negligence is considered in the bilateral case below.)

(b) The results about optimal care under the two rules (as well as most others in section 2) hold more generally than in the simple model considered above. They are valid if x is multidimensional and if x affects the probability distribution of harm (rather than just a single level of harm), as the reader can easily verify by essentially the argument given. Note that if a component of x is not observable by the court, the component cannot be included in the due care standard, so that the component would not be selected optimally under the negligence rule but would be under strict liability.

(c) No actual findings of liability occur under the negligence rule when due care is optimal, $\underline{x} = x^*$, since injurers are induced to take due care; but findings of liability do occur under strict liability (presuming that $p(x^*) > 0$).

2.2. Bilateral accidents and levels of care

Assume now that victims also choose a level of care y, that the probability of an accident is $p(x, y)$, which is declining in both variables, that the social goal is to minimize $x + y + p(x, y)h$, and that the optimal levels of care x^* and y^* are positive and unique.[3]

Under strict liability, it is evident that injurers' incentives are optimal conditional on victims' level of care, but victims choose $y = 0$; victims have no incentive to take care because they are fully compensated for their losses.

However, the natural version of the strict liability rule to consider in bilateral situations is strict liability with a defense of contributory negligence. Under this rule, a due care level $\underline{y}$ is established by courts for victims, and a victim is said to be contributorily negligent if his care level y was less than $\underline{y}$. An injurer is held liable for harm only if the victim was not contributorily negligent, that is, only if the victim's level of care was at least his due care level $\underline{y}$. If $\underline{y}$ is set by the courts to equal y^*, then it is a (Nash) equilibrium for both injurers and victims to act optimally. In particular,

[3] In some early, less formal literature on accidents, for example, Calabresi (1970), reference is made to the notion of the "least-cost avoider," the party—injurer or victim—who can avoid an accident at the lower cost. The idea of a least-cost avoider relies on the assumption that either party can undertake a discrete amount of care that is by itself sufficient to prevent an accident. Under this assumption, it is socially best for only the least cost-avoider to take care.

victims will choose y^* in order to avoid having to bear their losses, assuming that injurers choose x^*: victims obviously would not choose $y > y^*$. Also, victims would not choose $y < y^*$, for if they did so, they would bear their own losses (they would be contributorily negligent), so would minimize $y + p(x^*, y)h$; but for any $y < y^*$, this must exceed $y^* + p(x^*, y^*)h \geq y^*$. Conversely, injurers will choose x^*, assuming that victims choose y^*. For then injurers will have to pay for harm, so they will minimize $x + p(x, y^*)$.[4]

Under the negligence rule, if due care $\underline{x}$ is set equal to x^*, then injurers and victims will also act optimally in equilibrium. Injurers will choose x^* to avoid being liable, assuming that victims choose y^*. This is true by essentially the argument showing that injurers choose x^* under the negligence rule in the unilateral case. Victims will choose y^*, assuming the injurers choose x^*. This is so because victims will bear their losses (since injurers behave nonnegligently), meaning that they will select y to minimize $y + p(x^*, y)h$, implying that they will choose y^*.

A variant of the negligence rule is negligence with the defense of contributory negligence, according to which a negligent injurer is held liable only if the victim was not contributorily negligent. If due care levels are chosen optimally, $\underline{x} = x^*$ and $\underline{y} = y^*$, then injurers and victims will act optimally in equilibrium under this rule. The explanation is similar to that just given for the negligence rule.

Another version of the negligence rule is the comparative negligence rule. By definition of this rule, if both parties are negligent, they each bear a fraction of the harm, the fraction rising the lower their respective levels of care; if only one party is negligent, however, that party pays for the entire harm; and if neither is negligent, the victim bears his losses. Again, if $\underline{x} = x^*$ and $\underline{y} = y^*$, then injurers and victims act optimally in equilibrium under this rule. The explanation is identical to that under the negligence rule with the defense of contributory negligence. That the comparative negligence rule differs from the negligence rule when both parties are negligent is a moot aspect of the rule, since in equilibrium, the circumstance in which both parties are negligent is irrelevant to the calculations of either party.[5]

In summary, strict liability with the defense of contributory negligence and all of the versions of the negligence rules support optimal levels of care x^* and y^* in equilibrium, assuming that due care levels are chosen optimally. These conclusions were also established by Brown (1973).[6]

[4] It can also be shown that x^* and y^* is the unique equilibrium (when $\underline{y} = y^*$). Under the other rules to be discussed in this section, the equilibria are also unique.

[5] In versions of the model of accidents in which both injurers and victims might be found negligent in equilibrium, the comparative negligence rule obviously leads to different results from other versions of the negligence rule. See Bar-Gill and Ben-Shahar (2003), Edlin (1994), Cooter and Ulen (1986), and Rubinfeld (1987).

[6] Diamond (1974) proved closely related results shortly afterward. See also Green (1976), Emons (1990), and Emons and Sobel (1991), who focus on the case of heterogeneous injurers and victims.

It should be noted that courts need to be able to calculate optimal care levels for at least one party under any of the rules, and in general this requires knowledge of the function $p(x, y)$.

2.3. Unilateral accidents, levels of care, and levels of activity

Now let us reconsider unilateral accidents, allowing for injurers to choose their level of activity z, which is interpreted as the (continuously variable) number of times they engage in their activity. Let $b(z)$ be the injurer's benefit from the activity, where b is increasing and concave in z. Assume that $x + p(x)h$ is the cost of care incurred and the expected harm generated each time that an injurer engages in his activity, so that $z(x + p(x)h)$ is the total cost of care and expected harm given z.[7] Suppose that the social object is to maximize $b(z) - z(x + p(x)h)$, and let x^* and z^* be optimal values of x and z. Note that x^* minimizes $x + p(x)h$, so x^* is as described above in section 2.1. Thus, z^* is determined by $b'(z) = x^* + p(x^*)h$, which is to say, the marginal benefit from the activity equals the marginal social cost, comprising the sum of the cost of optimal care and expected accident losses (given optimal care).

Under strict liability, an injurer will choose both the optimal level of care x^* and the optimal level of activity z^*, as his objective is the same as the social objective, to maximize $b(z) - z(x + p(x)h)$, because damage payments equal h whenever harm occurs.

Under the negligence rule, an injurer will choose optimal care x^* as in section 2.1, but his level of activity z will be socially excessive. In particular, because an injurer will be led to escape liability for negligence by taking care of x^*, he will choose z to maximize $b(z) - zx^*$, so that z will satisfy $b'(z) = x^*$. Hence, if $p(x^*)h$ is positive, then $x^* < x^* + p(x^*)h$, so that concavity of b implies that $z > z^*$. The explanation for the excessive level of activity is that the injurer's cost of raising his level of activity is only his cost of care x^*, which is less than the social cost, as that also includes $p(x^*)h$. The excessive level of activity under the negligence rule will be more important the larger is expected harm $p(x^*)h$ from the activity.

The distinction between activity level and care level, and the result that under strict liability, both are chosen optimally, whereas under the negligence rule, the level of activity is excessive, is first developed in Shavell (1980b). Several comments about the conclusions about should be made.

(a) The failure of the negligence rule to control adequately the level of activity arises because negligence is defined here (and for the most part in reality) in terms of care x alone. A possible justification for this restriction in the definition of appropriate behavior is the difficulty courts would face in determining the optimal z^* and the actual z.

(b) The problem that the activity level is not properly controlled under the negligence rule has an analogue in respect to any component of behavior that would be difficult

[7] If the cost of care and expected harm do not rise linearly with z, the basic nature of the conclusions of this section would not be altered.

to incorporate directly into the negligence due care standard (either because it cannot readibly be observed (consider the frequency with which a driver checks his rear view mirror) or because it is not easy to calculate what constitutes the optimal setting of the component). Any such component of behavior will, however, be optimally controlled under strict liability.

2.4. Bilateral accidents, levels of care, and levels of activity

Suppose now that victims as well as injurers choose levels of care and of activity; let victims' level of activity be denoted t (victims' level of activity might be how many miles a pedestrian walks and exposes himself to risk) and victims' utility from it be $v(t)$, where v is increasing and concave in t. Suppose that social welfare is $b(z) + v(t) - zx - ty - ztp(x, y)h$, where, note, the assumption continues to be that expected accident losses rise linearly with (now victims' as well as injurers') level of activity. In this general situation, none of the liability rules that have been considered leads to full optimality. As just explained in section 2.3, the negligence rule leads injurers to engage excessively in their activity. Similarly, strict liability with a defense of contributory negligence leads victims to engage excessively in their activity (the number of miles pedestrians walk), as they do not bear their losses given that they take due care. Which form of liability, negligence or strict liability (with the defense of contributory negligence), is better will implicitly reflect whether it is more important to control victims' or injurers' level of activity; if injurers' level of activity is more important to control, strict liability will be superior, otherwise the negligence rule will be preferred.

The reason that full optimality cannot be achieved under either of the major types of liability rule is in essence that injurers must bear accident losses to induce them to choose the right level of their activity, but this means that victims will not choose the optimal level of their activity, and conversely. Indeed, for essentially this reason, it can be shown that there does not exist any liability rule that induces optimal behavior, x^*, y^*, s^*, and t^*, assuming that the liability rule is a function only of care levels x and y.[8]

3. Risk-bearing and insurance

Let us now consider the implications of risk aversion and the role of accident and liability insurance in relation to accident liability. For this purpose, it is convenient to consider the simple unilateral setting with injurers' level of care the only aspect of behavior at issue; how what is said will carry over to the more general context will be clear to the reader. Let U and V be the utility functions of injurers and victims respectively, assume that injurers and victims are either risk neutral or risk averse, and let u

[8] This result is shown in Shavell (1980b). However, fully optimal behavior can readily be induced with tools other than liability rules. For example, if injurers have to pay the state for harm caused and victims bear their own losses, both victims and injurers will choose levels of care and of activity optimally.

and v be their initial levels of wealth. The development below largely follows Shavell (1982a, 1987), which first studied liability, risk-bearing, and insurance.

3.1. First-best solution to the accident problem

The socially ideal solution to the accident problem is such that risk-averse parties do not bear risk and that the level of care is the optimal one, x^*, discussed in section 2.1.

In particular, the socially ideal solution to the accident problem can be identified with the solution that would be obtained by a dictator who could choose in a Pareto optimal way the level of care x and also levels of wealth contingent on the occurrence of accidents, subject to a resource constraint.[9] Denoting by v_n the wealth of a victim if he is not involved in an accident, v_a his wealth if he is, and similarly for u_n and u_a, the dictator would maximize

$$EV = (1 - p(x))V(v_n) + p(x)V(v_a)$$

subject to

$$EU = (1 - p(x))U(u_n) + p(x)U(u_a) = k,$$

where k is a constant, and subject also to

$$((1 - p(x))v_n + p(x)v_a) + ((1 - p(x))u_n + p(x)u_a) + p(x)h + x = u + v.$$

The second constraint is the resource constraint in expected value terms.[10] Is readily shown that when this problem is solved, (a) risk averse parties—be they injurers or victims—are left with the same level of wealth, and (b) the level of care is x^*, that minimizing $p(x)h + x$.

3.2. The accident problem given liability but in the absence of insurance

The question addressed here is what the Pareto optimal solution to the accident problem is in the presence of strict liability or the negligence rule. Formally, the problem is to maximize EV over the parameter(s) of the liability rule subject to the constraint that injurers choose care x to maximize EU under the liability rule, and subject also to the constraint that a lump sum ex ante transfer r be such that $EU = k$.

Under strict liability, the following can be shown. First, if injurers are risk neutral, the (Pareto) optimal magnitude of liability d is the harm h, and the first-best solution is achieved. The main reason is that, as injurers are risk-neutral, their bearing of risk

[9] In characterizing Pareto optimal solutions to the accident problem, we are of course characterizing social welfare optima. For were we to maximize any social welfare function depending positively on parties' expected utilities, the optimum would be Pareto optimal.

[10] A justification for writing the resource constraint in terms of expected values is that accident risks among the population are independent and that the population is large.

does not reduce their expected utility, and victims are protected against risk because, by definition of strict liability, they are compensated for harm. But second, if injurers are risk averse, the optimal magnitude of liability d is less than h and the first-best outcome is not achieved. The explanation is that, since injurers are risk-averse, it is desirable to lower liability a positive amount from h, since the first-order benefit in terms of reduce risk-bearing is positive, whereas the first-order loss from suboptimal incentives is zero or negative. Note that one interpretation of this result is that when injurers are risk-averse, it is not desirable to fully "internalize" a stochastic externality.

Under the negligence rule, the following is true. First, if victims are risk-neutral, the optimal due care standard $\underline{x}$ equals x^* and the first-best solution is achieved. This holds because injurers will be lead to choose x^* (if $d = h$) and thus will bear no risk (unlike under strict liability), and risk-bearing by victims will not reduce their expected utility as they are risk neutral. Second, if victims are risk-averse, then the optimal $\underline{x}$ is generally unequal to x^* and the first-best solution generally is not achieved. In this case, because victims are risk-averse, they bear risk when injurers take due care, implying that the first-best solution is not achieved, and also that it may be desirable for risk to be further lowered by raising $\underline{x}$ above x^*.

Note that consideration of risk-bearing alters the comparison of strict liability and negligence. Strict liability becomes more appealing when injurers are less risk averse than victims, since strict liability imposes risk on injurers. The negligence rule becomes more appealing when injurers are more risk-averse than victims, since the negligence rule imposes risk on victims.

3.3. The accident problem given liability in the presence of insurance

Now assume that victims can purchase insurance against accident losses that they might bear, and that injurers can purchase liability insurance against liability judgments that might be imposed on them. Assume also that the insurance policies are optimal for insureds in the usual sense that the policies are designed to maximize the expected utility of insureds, given the constraint that the insurance premium is actuarially fair (equal to expected coverage). Two assumptions about the information of insurers will be considered: that insurers can observe care x and thus make premiums depend on x; and that insurers cannot observe care (so that a situation of moral hazard exists).

In this case, then, the formal problem of finding the Pareto optimal liability rule is to maximize EV over the parameter(s) of the liability rule subject to the two constraints given in section 3.2 and subject also to the constraints that injurers and victims purchase insurance policies that maximize their expected utility given that premiums are fair.

Under strict liability, the following can be shown. The magnitude d of liability equals the harm h, and the first-best solution to the accident problem will be achieved unless injurers are risk-averse and liability insurers cannot observe care. Whether or not the first-best solution is achieved, it is not socially desirable for the government to interfere in the insurance market.

The foregoing is clearly true if injurers are risk-neutral, so let us consider the case where injurers are risk-averse. Then if liability insurers can observe care, it is straightforward that the first-best outcome will be achieved. The reason is that injurers will purchase full liability insurance coverage, will pay a premium of $p(x)h$, will choose care x to minimize the cost of care plus the insurance cost, namely, to minimize $x + p(x)h$, and hence will choose x^*. Since injurers do not bear risk (because coverage is full), victims do not bear risk, and care is optimal, the first-best outcome is achieved. The other case is that in which liability insurers cannot observe care, so that moral hazard exists. In this case, as a general matter, injurers will find it optimal (due to moral hazard) to purchase partial liability insurance coverage of $c < h$, so injurers will bear risk of $h - c$. Hence, they will have a positive incentive to take care, even though they own liability insurance. The first-best outcome will not be achieved because care will generally be different from x^* and because risk-averse injurers bear risk (of $h - c$). However, it can be shown that the outcome that results when damages $d = h$ and individuals purchase their privately-optimal liability insurance is second-best (equivalent to what a dictator could achieve if he was not able to directly control care but could control levels of wealth contingent on the occurrence of accidents). Hence, it can be demonstrated that government intervention in the liability insurance market is not desirable. This conclusion, about the undesirability of government intervention, is not transparent.[11]

Under the negligence rule, if due care $\underline{x}$ is set equal to x^*, the first-best solution is achieved. The explanation is, essentially, that injurers will be led to take due care and will not want to purchase liability insurance because they will not bear the risk of liability (it can be shown that they would not want to fail to take due care and buy liability insurance).[12] Further, since injurers take due care and are not held liable, victims are induced to purchase accident insurance so will not bear risk themselves.

Three points may be made in summary of the foregoing discussion. First, the presence of liability insurance affects incentives of injurers in a manner that depends on whether insurers can observe their level of care; liability insurance translates and modifies, but does not eliminate, injurers' incentives to take care under the threat of liability. Second, there is no basis for government intervention in liability insurance markets despite the moral hazard that liability insurance creates when insurers cannot observe care. This point is not only of intellectual note. Historically, the sale of liability insurance was viewed with skepticism and delayed in some countries, on the ground that it might interfere with incentives; in the former Soviet Union, liability insurance was proscribed;

[11] However, partial intuition for the conclusion may be helpful to provide. Suppose that it is assumed (rather than proved to be optimal) that the magnitude of liability is h. Then it must be welfare enhancing to allow injurers to purchase whatever liability insurance policy they please, for doing so must raise their expected utility, and this cannot affect the expected utility of victims since they are fully compensated for harm, given that liability is strict and that $d = h$.

[12] The conclusion that injurers do not purchase insurance under the negligence rule does not hold if courts might err in the negligence determination or if other uncertainties lead to the risk that injurers would be found liable. For a model of liability and insurance that considers this possibility, see Sarath (1991).

and today liability insurance is disallowed in this country in some domains.[13] Third, the presence of liability and accident insurance means that the risk-bearing reasons favoring either strict liability or the negligence rule, given above in section 3.2, are essentially mooted; if injurers are risk-averse, they can purchase liability insurance coverage under strict liability, and if victims are risk averse, they can purchase accident insurance coverage under the negligence rule.

4. Administrative costs

The administrative costs of the liability system are the legal expenses and the time and effort of litigants and of the state that are generated by the bringing and the resolution of suits. These costs are substantial; a number of estimates suggest that on average, administrative costs of a dollar or more are incurred for every dollar that a victim receives through the liability system.[14]

The factor of administrative costs raises the issue of whether the use of the liability system is worthwhile, that is, about the socially desirable volume of suits. The presence of administrative costs also leads us to reexamine optimal risk-reducing behavior and to inquire about the comparison between strict liability and negligence. In addressing these questions, it will be convenient to assume that parties are risk neutral and to consider the unilateral model of accidents with injurers' care being the only variable. The social goal will be taken to be minimization of social costs, comprised of the costs of care, expected harm, and now also expected litigation costs.

4.1. Volume of suit

In order to relate the private incentive to sue to what is socially desirable—and to show that they fundamentally diverge—it will be useful to examine in this section a discrete version of the unilateral model, where there is just a single positive level of care. (This allows us to isolate the issue of the volume of suit from the issue of the level of care, since care is not continuously variable.) Let x be the single level of care that injurers can exercise, p the probability of harm h if care is not taken, and $p' < p$ the probability of harm if care is taken. If $x < (p - p')h$, let us say that care is "efficient" since it lowers social costs, other things being equal. Further, let c_V be the cost of suit for a victim, c_I the cost of suit for an injurer, c_p the cost born by the public, and suppose that liability is strict.

When is suit socially desirable? In a general sense, the answer is: when the benefit of suit, in terms of inducing the exercise of care and reducing expected harm, exceeds

[13] On the historical resistance to the sale of liability insurance, see Tunc (1974, pp. 50–52). In this country today, liability insurance coverage is barred against punitive damages in some jurisdictions and also against certain types of fines; see Jerry (1996, pp. 471–477).

[14] See Danzon (1985, p. 187), Kakalik and Pace (1986, p. vii), and Tillinghast-Towers Perrin (2000, p. 12).

expected litigation costs. To amplify, if suit does not induce care to be taken, then plainly suit cannot be socially desirable: for if suit is not brought, social costs will be ph, whereas if suit is brought, social costs will be $p(h + c_V + c_I + c_P)$, which is higher due to litigation costs. Suppose now that suit does induce care to be taken. Then if suit is brought, social costs will be $p'(h+c_V+c_I+c_P)+x$. Hence, suit is socially worthwhile if and only if $p'(h+c_V+c_I+c_P)+x < ph$, or equivalently, if and only if the following key social condition holds:

$$p'(c_V + c_I + c_P) < (p - p')h - x.$$

This is exactly the condition that expected litigation costs are less than the net deterrence benefits of suit.

It is readily seen that suit may be brought by private parties even though that is socially undesirable. Suit will be brought by a victim of harm whenever $c_V < h$. This can be true even though suit is not desirable. For example, suit might not affect care at all—might not induce x—yet still be brought by victims since $c_V < h$; in this case, suit would be a pure waste, as the expected litigation costs of $p(c_V + c_I + c_P)$ would accomplish nothing. Even if suit induces care, the key social condition of the last paragraph for suit to be desirable might not hold when $c_V < h$. An aspect of this problem, note, is that in contemplating suit, the victim takes no account of the possibility that suit may fail to induce care or that the exercise of care may have little social value.

It is also evident that suit may not be brought even though it would be socially desirable that it be brought. This would be so if the bringing of suit would induce care and thereby produce a valuable deterrent; in other words the key social condition may hold even though $c_V > h$. An example is this: h is 100, c_V is 200, other litigation costs are 0, x is 1, p is 1, and p' is .01. Here suit is not brought, since 200 exceeds 100, so social costs are 100 (for p is 1). But were suit brought care of 1 would be exercised, the probability of harm would drop to .01, so expected social costs would be $.01(100 + 200) + 1 = 4$, which is far lower. The problem here is in part that when the victim considers suit, the fact that his willingness to sue would create very beneficial deterrence is of no moment to him.

Several remarks about the foregoing are worth making, helping to explain why the actual incentive to sue may be different from what is socially best. First, from the purely formal perspective, the private condition determining suit, whether $c_V < h$ holds, is facially quite different from the key social condition, $p'(c_V+c_I+c_P) < (p-p')h-x$. Second, there are two intuitively understandable social/private divergences at work. One is that the victim does not take into account the point that bringing suit causes a negative externality in the form of litigation costs on the injurer and society, namely c_I+c_P. The other is that, as mentioned, the victim does not take into account a positive externality due to suit, the creation of incentives to take care and to lower risks (for the victim sues only after an accident occurs). These two private/social divergences, working in opposite directions, make understandable how it is that there can be either too much or too little suit.

The private/social divergence may be of substantial importance in fact. For example, the financial reasons to sue for losses sustained in an automobile accident are often high and result in a tremendous amount of litigation and attendant expense—about half of all tort litigation in the United States concerns automobile accidents. Yet the incentive to drive safely may be relatively little affected by the prospect of suit (because incentives to drive safely, such as they are, have mainly to do with fear of injury from an accident or fear of criminal liability). If so, the large volume of automobile-accident litigation may largely constitute a social waste (my point is not so much that it *is* a social waste as it is that it could be; that the litigation is observed does not signal its social utility).[15]

The private/social divergence could be remedied by, for example, the state barring suit where it is undesirable but would be brought, or by the state subsidizing suit where it would be desirable but not be brought. For the state to do this, however, requires that it determine, among other things, the incentive that suit generates, in other words, that the key social condition must be ascertained. This imposes a high informational burden on the state. There is no easy fix, no simple remedy (such as making the victim pay the full litigation costs of suit) that will result in the socially desirable level of suit or necessarily to an improvement.

The points made here about the private versus the social incentive to use the legal system are first made in Shavell (1982b).

4.2. Level of care and volume of suit

Now let us return to the setting with care continuously variable, so that we can consider not only the issue of the volume of suit but also that of variation in the level of care. As will be seen, the level of care should be higher than the otherwise socially best level x^*, due to litigation costs. The reason is essentially that the occurrence of an accident creates greater social costs than just the harm h itself, as the social costs equal the harm plus litigation costs.[16]

To amplify, consider the following second-best problem: a dictator with the goal of minimizing total social costs can order victims when to sue and how much liable injurers should pay, but the dictator cannot directly control the level of care x. The solution to this second-best problem is a natural standard of comparison for the functioning of the liability system. It can be shown that the second-best solution has the following character. If suit is brought, injurers pay $h + c$, where $c = c_V + c_I + c_P$, so injurers should bear the harm plus total litigation costs. The explanation for this result is that, if suit is brought, then the occurrence of harm creates social costs of $h + c$, so that the injurer should be given an incentive to take care reflecting this amount. Additionally, under the second-best solution, suits are brought if and only if the following key social

[15] For empirical study of the effect of liability on automobile accidents, see Cummins et al. (2001), Dewees et al. (1996), E. Landes (1982), and Sloan (1998).

[16] This section follows Shavell (1999).

condition is satisfied:

$$p(x^*(h+c))c \leq [p(0) - p(x^*(h+c))]h - x^*(h+c),$$

where $x^*(h+c)$ is the x that minimizes $x + p(x)(h+c)$, that is, the optimal level of care when accident losses are $h+c$. The left side of this condition is expected litigation costs and the right side is the reduction in expected harm net of the cost of care induced by suit.

Under strict liability, with damages $d = h$, suit is brought when $c_V < h$. For reasons that are essentially those given in section 4.1, it is possible that suit is brought when the key social condition does not hold and suit is undesirable; and it is also possible that suit is not brought when the key social condition does hold and suit is desirable. If suit is brought, the level of care that will be taken is $x^*(h+c_I)$, since the injurer bears his own litigation costs; this level of care is less than the second-best optimum level of care of $x^*(h+c)$. In summary, suit might be excessive or inadequate; and when suit is brought, the level of care is too low.

The second-best outcome can be achieved, however, assuming that the state can determine whether the key social condition holds. For suit can be barred if that is optimal and subsidized if need be. Further, if suit is optimal and is brought, the correct level of care can be induced by raising damages by making the injurer pay $h + c_V + c_P$; since he naturally bears his own litigation costs, this level of damages means that the injurer's total expenses are $h + c$ so that his level of care will be second-best optimal.

4.3. Comments

(a) The basic points made above about the social versus private incentive to make use of the liability system, given that it is costly to employ, are robust. Whatever the specific nature of the model, it will tend to be true that a party contemplating making an expenditure will not take into personal account the negative externality that his expenditure will engender, in the form of expenditures by the other side of litigation and by the state; also, the party will not take into account the incentive effects of his expenditure on the behavior of others. As suggested in section 4.1, the possibility of a divergence seems to be of substantial policy interest because of the magnitude of administrative costs.

(b) Also, the point that the level of care (as distinguished from the volume of suit) should reflect the fact that when harm occurs and suit is brought, the social consequences are not limited to the harm but include litigation costs is general, and it yields a relatively easily applied policy prescription, that injurers' total payments should reflect total litigation costs. Such payments need not be in the form of damages; they could be in the form of fines on top of damages.[17]

(c) The difference between the private and the social motive to litigate is, as noted, initially developed in Shavell (1982b), and is extended in various ways in Menell (1983),

[17] An advantage of having litigation costs paid as fines is that then victims would not have an incentive to spend too much on litigation.

Kaplow (1986), Rose-Ackerman and Geistfeld (1987), and Spier (1997). Also, Polinsky and Rubinfeld (1988) consider the incentive to bring suit and injurers' level of care, presuming that the only policy instrument is the magnitude of damages. Under this assumption, the state is not able to induce both the optimal volume of suit and the optimal level of care, and as a consequence, the optimal level of damages does not equal harm plus others' litigation costs and may even be less than harm (if it is desirable to discourage suit and suit cannot be barred or taxed, damages need to be lowered). Polinsky and Che (1991), Hylton (1990), and Shavell (1999), however, allow for the level of suit to be controlled separately from injurers' level of care. For further discussion of litigation and private versus social incentives, see Shavell (1997) and Spier (2007).

4.4. Strict liability and negligence

Administrative costs enter into the comparison of strict liability and negligence as forms of liability. However, there does not seem to be a clear a priori difference in litigation costs under the two. Strict liability would be expected to result in a higher volume of cases than the negligence rule, for cases will be brought under strict liability whenever $c_V < h$, whereas cases will often not be brought under the negligence rule when $c_V < h$ since the injurer will often be known not to have acted negligently.[18] Although this factor of the volume of cases implies that strict liability is more expensive than the negligence rule, the litigation cost per case disputed is likely to be higher under the negligence rule, since negligence has to be determined, and since settlement of cases is less likely than under strict liability. Hence, either strict liability or the negligence rule could turn out to be the more costly on grounds of administrative costs, depending on the relative importance of the volume of cases (making strict liability more expensive) and the litigation cost per case (making the negligence rule more expensive).

Another issue bearing on strict liability versus the negligence rule concerns private and social incentives to bring suit. There is some basis for thinking that the problem of socially excessive suit discussed above is less likely under the negligence rule than under strict liability. To understand why, observe that if the negligence rule functions perfectly, it will be socially advantageous for suit to be subsidized so that it would be free to bring: for injurers will then decide to act nonnegligently, no suits will in fact be brought, and no litigation costs will be incurred. However, if as is realistic one assumes that courts may err in the negligence determination and/or that victims may not be able to tell whether injurers are nonnegligent, suits will sometimes be brought under the negligence rule. In consequence, the general qualitative results reached under strict liability will apply under the negligence rule as well, although the likelihood of excessive litigation would seem to be lower.

[18] Farber and White (1991) provide evidence that many medical malpractice cases are dropped when plaintiffs learn that the defendant probably was not negligent. Relatedly, Ordover (1978) examines a model in which victims' decision whether to sue under the negligence rule depends on their beliefs about the negligence of defendants.

Part B: Extensions

5. Liability of firms

Let us reconsider the theory of liability under the assumption that injurers are firms. This will require us to take into account the relationship between liability and market price, and also the level of production. Two cases will be distinguished: where accident victims are strangers to firms, such as where a person's car is damaged by a firm's truck; and where accident victims are the consumers of firms and are injured because of their purchase of a product or service, such as where a person's water heater ruptures and damages his property. Where accident victims are consumers, firms may be concerned about the risks generated by their products because consumers will pay less for risky products to the degree that they perceive risk; hence the need for liability as an incentive tends to be reduced.

Firms will be presumed to be identical, to maximize profits, and to operate in a perfectly competitive environment, so that product price will equal unit cost of production, including expected liability costs. Firms and victims will be assumed to be risk neutral, accidents to be unilateral (only firms' behavior affects risk), and the liability system to operate without administrative cost. The measure of social welfare is analogous to that considered in section 2 with activity levels: the utility consumers derive from products (such as from water heaters) and, where relevant, the utility that strangers obtain from their activities (such as from driving), minus expected harm, the costs of care, and direct costs of production. The development below follows the lines of Shavell (1987).

5.1. Victims are strangers

Let c be the direct production cost per unit of a firm's product, x the cost of care per unit of the product, s the quantity of the product produced and consumed, and $u(s)$ the utility consumers obtain from the product. Then social welfare is $u(s)-s[c+x+p(x)h]$, where the term in brackets is the production cost, cost of care, and expected harm suffered by strangers per unit. It is clear that optimal care x^* is, as earlier, the x minimizing $x + p(x)h$. Hence, the optimal quantity of the product s^* is determined by $u'(s) = c + x^* + p(x^*)h$, which has the familiar interpretation that marginal utility must equal the production cost of a unit, here including cost of care and expected harm.

Under strict liability the level of care and also the quantity produced will be socially optimal. Firms will choose x^*, in order to minimize unit production cost. Hence, unit production cost and price will equal $c + x^* + p(x^*)h$, implying that $u'(s) = c + x^* + p(x^*)h$, so that s will equal s^*. The quantity produced will be optimal because the price will reflect the expected harm caused by the product.

Under the negligence rule, if due care $\underline{x}$ equals x^*, then by the logic of section 2.1, firms will choose x^*, so that unit costs and product price will be $c + x^*$. Hence, the quantity sold will be determined by $u'(s) = c + x^*$, implying that the quantity under the negligence rule is socially excessive, since $c + x^* < c + x^* + p(x^*)h$. The reason is that the price does not reflect expected harm under the negligence rule. The problem with the negligence rule is analogous to that discussed in section 2.3; here the analog to the activity level is the quantity produced of the good.

Thus, the way that the liability rules function is similar to how they function when injurers are individuals; the effect of liability on firms' care is the same as on individual injurers' care, and the effect of liability on purchases via prices is analogous to its effect on individuals' choice of activity levels. The point that strict liability, but not the negligence rule, results in prices that induce optimal production is made in Polinsky (1980b) and Shavell (1980b).

5.2. *Victims are consumers*

Define the full price of the product as the market price plus the perceived expected accident losses that would be borne by a consumer. Consumers will choose their purchases s to maximize their utility $u(s)$ given the full price.

Suppose first that consumers can observe the risk $p(x)$ associated with a firm's product. Then the outcome will always be optimal. In the absence of liability, firms that choose x will charge $c+x$, and consumers will view the full price as $c+x+p(x)h$. Since consumers will buy from firms with the lowest full price, firms will be led to choose x^* to minimize the full price. Since the full price will be $c + x^* + p(x^*)h$, consumers will choose s to maximize $u(s) - s[c + x^* + p(x^*)h]$, so they will choose s^*. Under the negligence rule with due care of x^*, firms will choose x^* and charge $c + x^*$, the full price will again be $c + x^* + p(x^*)h$, so consumers will choose s^*. Under strict liability, firms will choose x^* to minimize the market price, which will be $c + x^* + p(x^*)h$; this will also be the full price (since consumers do not bear any losses), and consumers will choose s^*. That the presence of liability does not matter is due to the assumption that consumer information is perfect, meaning that market incentives result in proper care and consumer knowledge of the full price leads to correct purchases.

Suppose next that consumers cannot observe $p(x)$ for an individual firm but do know $p(x)$ on average. In the absence of liability, a firm would not choose positive x, for it could not charge a higher price for so doing. Hence, $x = 0$ in equilibrium, and since consumers know the risk to be $p(0)$ (as they know the average risk), the full price is $c + p(0)h$, and s satisfies $u'(s) = c + p(0)h$. Thus, care is too low and s is unequal to s^*, but s is optimal given $p(0)$. Under the negligence rule and under strict liability, firms act as they do in the previous case, so that x^* and s^* are chosen. In this case, then, the presence of liability matters, since care will not be taken without liability. Even the negligence rule results in the correct level of purchases s^* since consumers know the risk $p(x^*)$ in equilibrium.

Suppose now that consumers misperceive average risk, say they believe the risk is $\alpha p(x)$ where α is unequal to 1. To illustrate the difference this makes from the previous situation, suppose $\alpha < 1$, so consumers underestimate the average risk (the case of $\alpha > 1$ is analogous). Then in the absence of liability, firms again choose $x = 0$, and individuals choose s to minimize the full price of $c + \alpha p(0)h$, so they choose s

that is too high given the risk of $p(0)$. Under the negligence rule with due care of x^*, firms choose x^* and the full price is $c + x^* + \alpha p(x^*)h$, so that s is too high. Under strict liability, however, since firms choose x^* and the market price of $c + x^* + p(x^*)h$ is the full price, s^* is chosen. In other words, when consumers misperceive average risks, liability is needed to induce firms to take optimal care, and strict liability rather than the negligence rule is needed to induce consumers to purchase the correct quantity.

To summarize, when consumer information about product risk is perfect at the level of the individual firm, liability is not needed to induce optimal care nor to provide consumers with the right price on which to base purchases, since they take expected harm into account in deciding on purchases. When, however, consumer information is not perfect, liability is generally needed to achieve optimality, and the nature of the imperfection of information influences how, exactly, and the degree to which, liability improves outcomes.

Last, let us comment on warranties, which is to say, the effective choice of liability rules by consumers. Consumers will choose the form of warranty that results in the lowest full price. It follows that if consumers correctly perceive average risk, they will purchase warranties imposing either negligence or strict liability on firms, and the outcome will be optimal. If consumers misperceive risks, however, warranty choice may not lead to the optimal outcome. To illustrate, suppose that consumers underestimate average risk. Then consumers will tend not to want a warranty giving strict liability, for this would lead to a market price that appears expensive to them (in the extreme case where consumers believe the risk to be zero, the warranty would raise the price but appear to have no value to consumers). As a consequence, consumers might elect a negligence rule form of warranty, or none at all; thus, the outcome, would not be optimal; notably, purchases would be excessive. Therefore, misperception of risk may lead to problems with choice of warranties.

Literature on firms' liability to consumers, so-called product liability, includes early papers by Oi (1973) and Hamada (1976), assuming that consumers have perfect information about risk. Articles that study imperfect information about product risk and/or liability include Goldberg (1974), Schwartz and Wilde (1979), and Shavell (1980b). Key articles on the theory of warranties are Grossman (1981) and Spence (1977); see also Priest (1981) for an important informal treatment of warranties emphasizing consumer behavior.

6. Aspects of the negligence determination

In this section, I extend the simple model of the negligence determination in several ways. In doing this, I consider the unilateral model of accidents, but what is said about injurers will carry over to victims in regard to the determination of contributory negligence.

6.1. Differences among injurers

One issue that bears on the negligence determination concerns differences among injurers that lead to differences in the optimal level of care. In particular, suppose that injurers vary in their cost of exercising care (were the difference to concern some other factor, such as the likelihood of causing harm, the conclusions would be similar). Let kx be the cost of exercising care of x for an injurer of type k, where k is distributed on an interval $[0, K]$. The optimal level of care for an injurer of type k is the x that minimizes $kx + p(x)h$; denote this by $x^*(k)$, which is increasing in k.

If courts can determine an injurer's type, they can set the due care standard for each type optimally and induce all to act optimally; that is if $\underline{x}(k) = x^*(k)$, then each type of injurer will take due care and act optimally. However, if courts cannot observe k, and must set a single due care standard $\underline{x}$ in $[x^*(0), x^*(K)]$, then it can be shown that the optimal level of this uniform due care standard is greater than or equal to $x^*(E(k))$, the optimal level of care for the person with the mean cost parameter. (The optimal level of care for this mean person is sometimes referred to as the standard of due care for the "reasonable" man.)

Now make the additional assumption that injurers have a choice whether or not to engage in their activity, and if they so do they obtain a benefit of u (there is only one level of activity, for simplicity). Then it is socially beneficial for an injurer of type k to engage in the activity if and only if $u > kx^*(k) + p(x^*(k))h$. However, we know from above (see section 2.3) that, under the negligence rule, injurers may decide to engage in the activity even though doing so is socially undesirable; the problem is that injurers can avoid having to pay for the activity by taking due care. This problem is most serious in regard to injurers with high k—for whom it is very expensive to take care or who are awkward or inept; for them $kx^*(k) + p(x^*(k))h$ will be high—perhaps greatly exceeding u—since $x^*(k)$ will be low. This has an implication for the setting of due care. Even if the courts can observe injurers' type k, the optimal level of due care may not be $x^*(k)$ for all k; rather optimal due care may be $\underline{x} > x^*(k)$ for all k above some threshold k'. By setting such a due care standard, engaging in the activity may become too expensive to be worthwhile for high k types, thus implicitly combating the problem of excessive engagement in the activity for these most dangerous types. For similar reasons, in the case where k cannot be observed, the optimal single level of due care will tend to be higher than in the last paragraph.[19]

6.2. Imperfect assessment of care by courts

Another issue that bears on the determination of negligence is that courts may make mistakes in assessing the care exercised by injurers. Let e be the error in the level of

[19] On differences among individuals and the negligence determination, see Diamond (1974) and Shavell (1987, pp. 86–91).

care, so that $x + e$ is the level of care observed by a court. Thus, if $\underline{x}$ is the due care standard, negligence will be found when $x + e < \underline{x}$, which is to say, when $e < \underline{x} - x$. Accordingly, even if $x = \underline{x}$, negligence might be found in the event of an accident, for if $e < 0$ negligence would be found.

An injurer subject to the negligence rule will choose x to minimize $x + \text{Prob}[x + e < \underline{x}]p(x)h$, where Prob denotes probability. It can be shown that if the due care level $\underline{x}$ would be met in the absence of error, then in the presence of error injurers will choose $x > \underline{x}$, provided that the distribution of e is sufficiently concentrated. In particular, care will be excessive under this condition if due care $\underline{x}$ is set at the optimal level x^*. Also, negligence will sometimes be found in general. The intuition behind this result that care tends to be excessive due to error in its determination is that the injurer can guard against being found negligent by taking excessive care; for taking more than due care means that even if actual care is underestimated, he might still escape liability, and taking more than due care may well cost less than the reduction in expected liability that it brings about. The analysis of error in observing care is first examined in Diamond (1974) and the point about excessive care is emphasized in Craswell and Calfee (1986).[20]

6.3. Imperfect ability to control care by injurers

An additional issue that affects the finding of negligence is that individuals may not be able to control perfectly what might be called their momentary level of care, such as a driver's behavior at a particular instant. A driver might be unable to control his or her behavior at a particular instant due to a sneeze (causing the driver to swerve) or to an inadvertent lapse in attention. In contrast, what an injurer is able to control is his general habit of attention, his mental attitude (which might be reflected, for instance, by the frequency with which a driver looks at his rear view mirror, how frequently he switches lanes, and the like). But a court would usually find it difficult to observe this mental attitude, as opposed to the momentary level of care. We might formalize these observations by assuming that an individual chooses x, his general level of attention, whereas his momentary level of care is $x + e$ where e is a random term, and where it is $x + e$ that the court observes. Hence, the injurer will be found negligent in the event of an accident if $x + e < \underline{x}$, and the analysis of his behavior will be similar to that in the last section, with the conclusion being that care will tend to be excessive if $\underline{x} = x^*$, and that negligence will sometimes be found.[21]

6.4. Errors in the calculation of due care

A further issue concerns whether courts may make mistakes in ascertaining the optimal due care standard. If courts' information about the function $p(x)$ or about the cost of care is imperfect, then their calculation of x^* will tend to be incorrect.

[20] The analysis is further developed in Shavell (1987, pp. 93–99).

[21] On the momentary level of care, see Diamond (1974) and Shavell (1987, pp. 96–97). See also Grady (1983).

One possibility is that injurers do not know in advance how the courts will err in determining x^* and thus in determining due care $\underline{x}$. In this case, injurers will tend to take excessive care in order to escape liability due to mistake, and negligence will sometimes be found, for reasons closely associated with those discussed in the last two sections.

The other possibility is that injurers anticipate what the (mistaken) due care standard will be. Under this assumption, if due care is too low, $\underline{x} < x^*$, injurers will take care of $\underline{x}$ and will not be found negligent. If due care is too high, $\underline{x} > x^*$, injurers will take care of $\underline{x}$ and will not be found negligent, provided that $\underline{x}$ does not exceed x^* by too much. If $\underline{x}$ is sufficiently high, however, injurers will decide to be negligent and will choose x^*.[22]

7. Causation

A fundamental principle of liability law is that a party cannot be held liable unless he was the cause of losses. For example, if a surgeon negligently performed a procedure and his patient died on the operating table, but the patient would have died even if the surgeon had properly carried out the procedure, the surgeon would not be held liable. Or if a firm pollutes a stream with a carcinogenic agent and a person living nearby develops cancer, but the cancer is shown to be due to another carcinogenic factor, the firm will not be held liable.

The idea of causation of harm can be expressed as follows. Let s denote a state of the world in the universe of possible states. Let $h(s, x)$ be the harm that occurs in state s if action x (which could be a level of care) is taken. Then, given a state s, taking action x may be said to be a cause of losses of h relative to some other reference or comparison action x_0 if $h(s, x) - h(s, x_0) = h$. Often, when it is said that an injurer is a cause of losses, the comparison action is not mentioned but it will be implicit from the context.

Let us now consider the effect of the causation requirement—that liability is imposed only if the injurer was the cause of harm—on incentives under liability rules. On this issue, see originally Calabresi (1975) and Shavell (1980a).[23]

7.1. *Strict liability*

Under strict liability, it is readily shown that if injurers are held liable for the harm that their activity causes, but not for other harm, then their incentives to take care and to engage in harmful activities will be optimal. Here, by harm that their activity causes is meant harm that is caused by their engaging in their activity relative to their not engaging in their activity.

If injurers are held liable for less than the harm they cause, or for more than that harm, their incentives will generally be suboptimal. Notably, if they were held liable for harm that they do not cause—if firms are held liable for cancer that their pollution does not

[22] See Shavell (1987, pp. 97–98).

[23] See also W.M. Landes and Posner (1987, pp. 228–255) and Shavell (1987, pp. 118–123).

cause—then they might be led to take excessive care[24] or be undesirably discouraged from engaging in their activity.

7.2. Negligence rule

Under the negligence rule, it can be shown that injurers will be led to take due care, assuming that due care equals optimal care, if they are held liable only if they caused harm and pay damages equal to harm. Here two possible definitions of harm are natural to consider. One is harm relative to the harm that would have occurred had the injurer not engaged in his activity (for example, not driven a car). The other possible definition is harm relative to the harm that would have occurred had the injurer taken due care; in other words, harm would be measured by $h(s, x) - h(s, \underline{x})$. I have employed the first definition in the present chapter and will continue to do so below.[25]

If injurers were held liable more often than when they caused harm, this would not lead to their taking excessive care; it would only augment their already sufficient incentive to take due care. Thus, unlike under strict liability, there is not an affirmative reason to restrict liability to accidents caused by injurers under the negligence rule in the simple unilateral model of accidents.

Nevertheless, there exist advantages of restricting liability on the basis of causation under variations of the basic unilateral model. Specifically, in the presence of uncertainty about the negligence determination, relaxation of the causation requirement might exacerbate the tendency to take excessive care (this tendency exists under the assumption that damages are measured according to the first definition of harm). Further, relaxation of the causation requirement would tend to increase the volume of litigation and thus administrative costs.

7.3. Uncertainty over causation[26]

In many situations there is uncertainty about causation. For example, it may not be known which manufacturer out of many sold the product that resulted in injury, or whether harm was due to the defendant firm's pollution or to background factors. The traditional approach of the law is to employ a 50% probability threshold: to hold a defendant liable if and only if the probability that the defendant was the cause of losses

[24] To illustrate why an injurer might take excessive care, imagine that a firm would be made to pay for cancers that its pollution did not cause if it discharges some pollution, but not otherwise. Then the firm would have an excessive incentive to reduce pollution to zero (perhaps by installing a socially undesirably expensive pollution control device).

[25] The main reason that I adopt the first definition of harm is that courts frequently have insufficient information about $h(s, \underline{x})$ to calculate $h(s, x) - h(s, \underline{x})$. Were the second definition of harm employed, it can be shown that injurers would still be induced to take care of x^* if $\underline{x} = x^*$. The main differences that the second definition of harm would make to the analysis flow from the implication that expected liability would be continuous in x at $\underline{x}$ rather than discontinuous. Kahan (1989) and Grady (1983) focus on the implications of the second definition of harm.

[26] On the subject of this section, see Rosenberg (1984) and Shavell (1985), and Shavell (1987, pp. 123–126).

exceeds 50%. More generally, the courts could employ a probability threshold t different from 50%.

The probability threshold approach may lead to suboptimal behavior. On one hand, if an injurer knows that the probability that he was the cause of losses will be less than 50% (or whatever is t), he will never be held liable so will have no motive to reduce risk. On the other hand, if he anticipates that the probability that he was the cause of losses will exceed 50%, he will always be held liable, and his incentives will also be suboptimal, at least under strict liability. For under strict liability, his excessive liability implies that he will have too low an incentive to engage in the activity. Under the negligence rule, however, his being held liable even when he was not the cause of losses will only increase his incentive to take due care, so should not result in excessive care, assuming that the negligence determination occurs without error.

As an alternative to the probability threshold criterion, the legal system has sometimes adopted the proportional liability approach, that is, of imposing liability in proportion to the likelihood of causation. Under this approach, an injurer whose likelihood of being the cause of harm was q would pay damages of qh in a case for which the harm was h. (The proportional approach is often called market-share liability, since in a context in which different firms cause accidents and it is hard to identify which firm caused a particular harm, the likelihood of causation may be approximated by a firm's market share.) The proportional liability approach may lead to the same incentives to reduce risk that exist in the absence of uncertainty over causation. In particular, suppose that $p(x)$ is the probability that an injurer would cause an accident, and that c is the probability of an accident due to other causes (assume these are mutually exclusive events for simplicity). Then, the probability of harm is $c + p(x)$ and the probability q that harm was due to the injurer, as calculated by the court, will be $q = p(x)/(c + p(x))$. (Note that the court must know $p(x)$ in order to compute q.) Under strict liability using the proportional approach, the probability of having to pay damages is $c + p(x)$ and the amount paid would be qh, so that expected liability would be $(c + p(x))qh = (c + p(x))[p(x)/(c + p(x))]h = p(x)h$, implying that care, x, will be chosen optimally and the decision to engage in the activity will also be optimal.[27] Likewise, incentives to take due care under the negligence rule will be correct if due care is set optimally.

7.4. Proximate causation[28]

Even if a party is a cause of losses, he may still escape liability under the law because he was not the "proximate cause" of losses. There are two major categories of losses that

[27] Incentives will not be optimal if, however, the courts cannot properly calculate $q = p(x)/(c + p(x))$. A context in which problems might arise is where many firms market a product that causes harm and market share is employed to determine q. It might be that a particular firm's product is safer than another's (one firm's drug might cause fewer side effects than another's). If so, its $p(x)$ is lower than another's; but if only market share, and not $p(x)$, is considered by courts, incentives would not be optimal.

[28] For analysis of proximate causation, see Shavell (1980a) and Shavell (1987, pp. 110–115, 121–123).

are said not to be proximately caused. One is accidents that came about in an atypical, freak manner, for example, where a dog imbibes nitroglycerin left at a mining site and then explodes, injuring nearby persons. Allowing parties to escape liability for unusual accidents is sometimes thought not to undermine incentives, on the ground that no one could have foreseen such accidents. This argument, however, is subject to the criticism that courts may find it difficult to discriminate between accidents that can and cannot be foreseen and that it reduces the incentives of injurers to become informed. Moreover, the argument leads to the reductio ad absurdum that there should never be liability: after all, any accident may be viewed as extraordinarily unlikely (of essentially zero probability) if it is described in sufficient detail.

The other major category of accidents said not to be proximately caused, and for which parties are not held liable, are those coming about on account of what might be described as coincidence. For example, in a well known case, a speeding bus happened to be at just the "right" point on its route to be struck by a falling tree. Here, the bus company escaped liability for the injuries to passengers even though the injuries would not have occurred but for the excessive speed of the bus. Allowing parties to escape liability for such coincidental accidents might not affect incentives to take precautions, however. If the probability of a bus being struck by a falling tree is independent of its speed, failing to impose liability when trees fall on buses would not affect incentives to speed.

8. The magnitude of liability: damages

8.1. Basic theory: strict liability

Under strict liability, if the magnitude of liability, which as noted is called damages d, equals harm, then incentives to take care will be optimal in the unilateral model of accidents. For if so, the injurer's problem will be to minimize $x + p(x)h$, which is the social problem, so he will choose x^*; whereas if $d < h$ the level of care will be inadequate and if $d > h$ the level of care will be excessive. Likewise, for levels of activity to be optimal, damages must equal harm.[29]

The point that damages equal to harm is optimal essentially carries over to the situation, not yet considered, where the magnitude of harm is stochastic. Suppose that, if an accident occurs, the harm h is governed by a probability distribution with density $f(h;x)$; thus, note, the level of care may affect the probability distribution of damages. The social problem in this situation is to minimize $x + p(x)\int hf(h;x)dh = x + p(x)E(h;x)$, where $E(h,x)$ is expected harm conditional on an accident occurring

[29] In the bilateral model of accidents, damages equal to harm is optimal under the rule of strict liability with a defense of contributory negligence, if victims' activity level is taken to be fixed. If, however, victims' activity level is variable, then optimal damages may well be less than harm, for then some part of losses will be borne by victims and will induce them desirably to moderate their level of activity.

when care is x. It is clear from the expression of the social objective that if $d = h$, the injurer will minimize the social objective so will choose x^*.

Several observations are worth making about this more general situation where the magnitude of harm is stochastic.

(a) It is not necessary that $d = h$ for incentives to be correct. If damages equal $E(h; x)$ regardless of the actual h, then incentives will be correct.

(b) Legal doctrines that impose a ceiling on damages (on the ground that no one could have expected such high damages) or that exclude damages when the probability density f is sufficiently low (on the ground that no one could have expected such an unlikely event) lead to inadequate care and excessive activity levels; for such doctrines truncate expected liability conditional on an accident occurring to a level below $E(h; x)$.

(c) Legal doctrines that impose damages exceeding harm for certain types of outcome (see section 8.4 below on punitive damages) lead to excessive care and inadequate activity levels.

8.2. Basic theory: negligence rule

Under the negligence rule, analysis of the optimal magnitude of damages is different: on one hand, damages can be somewhat less than harm and optimal care will still be induced; on the other hand, damages can exceed harm and optimal care will be induced. To begin with, consider the simple model with one level of harm h, due care $\underline{x}$ set optimally at x^*, and no errors in the negligence determination. It was shown in section 2.1 that if $d = h$, the injurer will take due care of x^*. In fact, the incentive to take care of x^* rather than less care is sharp, in the sense that expected liability rises discontinuously from 0 to a positive level if x is reduced from x^*, since then expected liability jumps from 0 to $p(x)h$. This discontinuity in expected liability can be shown to imply there is an interval $[h, h - k]$ for some positive k such that if d is in the interval, the injurer will still be induced to choose x^*. If $d > h$, then it is obvious that x^* will be taken and that greater care will not be taken, since if care is just x^* then the injurer will be relieved of liability.

These points require some qualification, however.

(a) If there is uncertainty in the negligence determination, then damages exceeding harm may exacerbate the problem of excessive care (see section 6.2).[30] In this situation, damages equal to harm will not induce optimal care, but some level of damages (perhaps less than h) will induce optimal care.

(b) If the magnitude of harm is stochastic, the points made in the preceding paragraph largely carry through.

[30] An additional issue is that erroneous findings of liability tend to remedy the problem of excessive levels of activity under the negligence rule, raising the possibility that setting damages above h would be desirable.

8.3. Difficulty in estimating harm

Some components of harm are hard to estimate, for example, the decline in profits that will be caused by a fire at a store. The courts sometimes exclude such difficult-to-measure elements of harm from damages (on the ground that they are too speculative). This legal practice might be justified if the cost of ascertaining a component of harm outweighs the value of the improvement in incentives that its inclusion would accomplish. However, the cost of estimating a component of loss would be low if rough estimates are used, suggesting that that approach may be superior to omitting speculative components of harm from damages.

8.4. Punitive damages

When an injurer's behavior departs substantially from what is appropriate, damages in excess of harm, so-called punitive damages, may be imposed. If imposition of such damages causes expected liability to exceed expected harm, injurers will be induced to take excessive precautions, at least under strict liability, and they will also reduce their levels of activity undesirably. Under the negligence rule, excessive damages may also lead to undesirable levels of care when there is uncertainty about the negligence determination, as discussed in section 6.2. Hence, the use of punitive damages seems socially detrimental as a general matter.

There is, however, a fairly general circumstance in which damages exceeding liability are desirable: when injurers sometimes escape liability. This possibility may arise because injurers are hard to identify as the sources of harm (the origin of pollution may be difficult to trace) or because victims do not choose to bring suit (litigation costs may discourage legal action). If injurers who ought to be found liable for harm h are in fact only found liable and made to pay damages with probability q, then if damages are raised to $(1/q)h$, injurers' expected liability will be h. Thus, the more likely a party is to escape liability, the higher should be damages when the party is found liable. Accordingly, a firm that dumps toxic wastes at night, or an individual who tries to conceal a bad act, should have to pay punitive damages, but not an injurer who causes harm in a noticeable way. On these points and others, see, for example, Cooter (1989), Diamond (2002), and Polinsky and Shavell (1998).[31]

Another assumption that would justify damages exceeding harm is that injurers obtain "illicit" utility—utility not credited in social welfare—from certain harmful acts, such as spitefully destroying a person's property. Under this assumption, it is readily shown that damages exceeding harm may be desirable, for they are required to induce injurers not to commit the harmful acts, in order to offset their illicit utility. The assumption of illicit utility might be considered problematic, however, as it implies that social welfare

[31] See also Daughety and Reinganum (1997). For empirical study of punitive damages, see Eisenberg et al. (1997), Karpoff and Lott (1999), and Polinsky (1997).

is not a function of individuals' utilities (and thus leads to conflicts with the Pareto principle[32]).

8.5. Accuracy of damages and legal costs

Much expense is incurred in litigation about the magnitude of a victim's harm, which raises the question of what the social value of greater accuracy is and whether the private value of accuracy is different from the social value. It turns out that the private value of information about harm tends to exceed the social value. To explain, there is social value in establishing harm accurately primarily when injurers know, at the time that they choose their level of care, how much harm they might cause, for then they can take appropriate steps to prevent harm. For example, if an injurer anticipates that the atypically large harm he might cause will be accurately measured, he will exercise a properly high degree of care, as is socially desirable. However, injurers often lack considerable information about the harm they might cause when they decide on their precautions. Drivers, for example, know relatively little about how much harm a potential victim would suffer in an accident (the seriousness of injuries, the magnitude of lost earnings). In such cases, there is limited social value in assessing damages accurately because injurers behavior would not be much affected by doing so; drivers' incentives to avoid accidents would be largely the same if, instead of striving for fairly precise measurements of harm, courts employed rough averages (based, perhaps, upon abbreviated litigation over damages or upon figures from a table).[33] Regardless of the possibly attenuated social value of accuracy in the measure of damages, victims and injurers have very strong private incentives to spend to establish damages accurately in court. A victim will always be willing to spend up to a dollar to prove that harm is a dollar higher, and an injurer will always be willing to spend up to a dollar to prove that harm is a dollar lower. This suggests why, as shown in Kaplow and Shavell (1996b) the private value of information about damages exceeds the social, so that too much is spent by litigants proving damages.

9. Multiple injurers

In some contexts more than one injurer jointly contributes to harmful events, such as when two drivers race each other and cause another car to drive off the road, or where two firms operate a refinery which explodes. For simplicity, let us consider the unilateral model of accidents and that there are just two injurers (the case of n injurers is essentially identical). Let $p(x_1, x_2)$ be the probability of harm, so that the social problem is

[32] See Kaplow and Shavell (2001).

[33] A qualification to this point, emphasized by Spier (1994), arises where the probability distribution of harm is affected by an injurer's degree of care. If this is so, then accuracy in assessing harm will influence an injurer's incentives to reduce risk even when, at the time he chooses his level of care, he does not have information about the harm that would occur in an accident.

to minimize $x_1 + x_2 + p(x_1, x_2)h$, and let x_i^* denote the optimal x_i, which are assumed to be positive.

9.1. *Strict liability*

By strict liability is meant imposition of damages on injurers in amounts that are independent of their levels of care and that the sum of damages equals h (the usual rule).

Suppose that the injurers act independently of one another. Then the optimal outcome is not achievable, since to induce optimal care of x_i^* injurer i must bear damages of h under strict liability.

If injurers act in concert, then it is obvious that they will behave optimally since their problem will be the social problem.

9.2. *Negligence rule*

By the negligence rule is meant the rule under which each injurer escapes liability if his care is at least his due care level $\underline{x}_i$; if he alone is negligent, he bears damages of h, and if he and the other injurer are both negligent, they will jointly pay h, in amounts that may depend on their x_i.

Suppose that injurers act independently of each other. Then, if due care levels are optimal, $\underline{x}_i = x_i^*$, it is clear (from the argument in section 2.1) that each injurer taking due care is an equilibrium, and it can be shown that it is the unique equilibrium. Thus the negligence rule is superior to strict liability. The superiority of the negligence rule has to do with the fact that, by the nature of the negligence rule, each party is threatened with damages equal to the *entire* harm h when other parties act optimally; whereas under strict liability, each party is threatened with a lesser amount.

If injurers act in concert, then they will also act optimally if due care levels are optimal.

These points about behavior of multiple injurers under the negligence rule were first made by W.M. Landes and Posner (1980).

10. The judgment-proof problem

The possibility that injurers may not be able to pay in full for the harm they cause is known as the judgment-proof problem and is of substantial importance, for individuals and firms often pose risks significantly exceeding their assets (a person of modest means could cause a devastating fire; a small firm's product could cause many deaths). The judgment-proof problem reduces the effectiveness of accident liability in combating risk and also lowers the incentive to purchase liability insurance. On the judgment-proof problem, see originally Summers (1983) and Shavell (1986).

10.1. Incentives to engage in harmful activities

When injurers are unable to pay fully for the harm they may cause, their incentives to engage in risky activities will be greater than otherwise. Under strict liability, their incentives to engage in an activity are optimal if their assets are enough to cover the harm they might cause, but their incentives will be inadequate if they are unable to pay for the harm. Under the negligence rule, the situation is somewhat different. If the rule functions without error and injurers are induced to take care (as they might despite being judgment proof—see below), they bear no liability and their incentives to engage in the activity (which are too great, by the argument of section 2.3) are not altered by the existence of the judgment-proof problem. If, though, injurers are not induced to take optimal care, or there are errors in the negligence determination that sometimes result in findings of negligence, then the existence of the judgment-proof problem does lead injurers to engage more often in the activity than they otherwise would.

10.2. Incentives to take care

Likewise, when injurers are unable to pay for all the harm they might cause, their incentives to take care tend to be diluted. Consider the injurer's problem of choosing care x under strict liability, when his assets are $y < h$. The injurer's problem is often formulated as minimizing $x + p(x)y$; hence, the injurer chooses $x(y)$ determined by $-p'(x)y = 1$ instead of $-p'(x)h = 1$, so that $x(y) < x^*$ (and the lower is y, the lower is $x(y)$). Under this formulation of the injurer's problem, note that it is implicitly assumed that care is nonmonetary, for if care is monetary, the injurer's wealth after spending on care would be $y - x$, and only this amount would be left to be paid in a judgment. If care is monetary, in other words, the injurer's problem is to minimize $x + p(x)(y - x)$. The solution $x(y)$ to this problem is determined by the condition $-p'(x)(y - x) = 1 - p(x)$. The interpretation is that the marginal benefit of taking care, saving the person's remaining wealth of $y - x$ with a higher probability of $-p'(x)$, equals the marginal cost of the expenditure, which is not 1 but $1 - p(x)$. In effect, spending on care has a private cost of less than the social cost of 1, since there is a chance $p(x)$ that an expenditure will not be a private cost as it would be paid in a judgment. This gives rise to the theoretical possibility that care could be excessive, $x(y) > x^*$; but that would not be the case when p is low. For simplicity, I will assume below that care is nonmonetary.

Under the negligence rule, with due care $\underline{x}$ set equal to x^*, the injurer might choose x^* rather than risk liability with a lower x and incur costs of $x + p(x)y$; indeed, since $x^* + p(x^*)h < x^*$, there exists an interval $[h - k, h]$ with positive k such that if y is in this interval, x^* will be induced. So, unlike with strict liability, the judgment-proof problem does not lead to inadequate care under the negligence rule unless assets y are sufficiently less than harm h.

10.3. Incentives to purchase liability insurance

Risk-averse injurers who may not be able to pay for the entire harm they cause will tend not to purchase full liability insurance or any at all. This is because purchase of full coverage would require them to pay premiums for damage payments that they would not make in the absence of coverage: if a person with assets of $10,000 buys coverage against liability of $100,000, he is purchasing coverage against $90,000 of losses that he would not suffer if he did not have coverage.[34] The nature and consequences of this effect of the judgment-proof problem depend, among other things, on whether liability insurers have information about risk and thus link premiums to it. If liability insurers have information about risk and base premiums on it, then reduction in the purchase of liability insurance tends to reduce incentives to take care (as in the last section). If liability insurers do not have information about risk, then reduced purchase of liability insurance tends to increase incentives to take care (even though the incentives generally remain suboptimal). In both cases, reduction in the purchase of liability insurance tends to undesirably increase incentives to engage in the harmful activity.

10.4. Policies to ameliorate the judgment-proof problem

Several types of policy responses to the dilution of incentives caused by the judgment-proof problem are of interest. If there is a second party who has some control over the behavior of the party whose assets are limited and who directly affects risk, then the second party can be held vicariously liable for the losses caused by the first, inducing the vicariously liable party to cause risks to be lowered. Thus, holding a large contractor liable for the accidents caused by a small subcontractor or an employer for accidents caused by its employees will lead the former to control the risks posed by the latter. See section 11.

Additionally, parties with assets less than a specified amount could in some contexts be prevented from engaging in an activity. However, such minimum asset requirements may undesirably prevent some individuals who ought to engage in the activity from doing so; although their assets are low and their care would be inadequate, their benefits might still exceed the expected harm that they create. Minimum asset requirements are a somewhat blunt instrument for alleviating the incentive problems under consideration.[35]

A third response to inadequate incentives caused by the judgment-proof problem is regulation of liability insurance. One form of insurance regulation would mandate purchase of (perhaps full) liability insurance coverage. This approach would be especially appealing when insurers can observe the precautions taken by injurers and link premiums to it. In that case, if injurers purchase full liability insurance coverage, their incentives to reduce risk would be optimal (see section 3.3). However, if insurers cannot

[34] See Huberman et al. (1983), Keeton and Kwerel (1984), and Shavell (1986).

[35] On minimum asset requirements, see Shavell (2005).

observe risk and moral hazard exists, then requiring the purchase of liability insurance further reduces the incentive to take care. Therefore, mandating the purchase of liability coverage may not be desirable. This point suggests that an opposite form of insurance regulation may be advantageous: barring purchase of liability insurance (a policy that is sometimes employed, as noted in section 3.3). Forbidding purchase of liability insurance can improve incentives to take care if, without a prohibition, injurers would have purchased positive coverage and insurers cannot observe injurers' level of care.[36]

Fourth, the use of corrective taxes equal to expected harm may help to alleviate the judgment-proof problem. When harm is caused with a low probability, the expected harm is much less than the actual harm. Hence, parties with limited assets may be able to pay the appropriate tax on risk-creating behavior even though they could not pay for the harm itself. Suppose that the release of pollution will cause harm of $1 million with a 1% probability. Thus, a firm with $100,000 of assets would be able to pay $10,000 for the expected harm it would cause were it to cause the pollution, even though it would only be able to pay one tenth of the actual $1 million harm it might generate, and so its incentives to reduce risk would be much too low under the liability system.

A fifth way of correcting for dilution of incentives is for the state to regulate parties' behavior directly, such as with traffic laws or safety requirements. Regulation, however, may involve inefficiency because of regulators' limited knowledge of risk and of the cost and ability to reduce it.

Another way of mitigating dilution of incentives is resort to criminal liability. A party who would not take care if only his assets were at stake might be induced to do so for fear of imprisonment.

11. Vicarious liability

Vicarious liability, as mentioned in the last section, is the imposition of liability on a party related to the actual author of harm, where the vicariously liable party usually has some control over the party who directly causes the harm. The forms of vicarious liability vary widely and include holding parents liable for harms caused by their children, contractors for harms caused by subcontractors, firms for the harms caused by employees (a particularly important example), and even, occasionally, lenders for the harms caused by borrowers.

There are two major reasons that vicarious liability may be socially desirable. One is that the injurer may not have proper information about reduction of harm, whereas the vicariously liable party may have good, or at least superior, information and be able to influence the risk-reducing behavior of the injurer. This feature of vicarious liability might apply, for example, in regard to a parent in relation to a child (suppose a teenage child does not realize how dangerous using a motorboat might be to swimmers and

[36] On regulation of liability insurance, see Jost (1996), Polborn (1998), and Shavell (1987, 2000, 2005).

that the parent can monitor the child's use of the boat), or in regard to a firm and its employees (suppose a worker does not understand how dangerous a toxic waste product is, but the firm does and thus controls how it is transported by the worker to a dump site).

The other basic reason that vicarious liability may be desirable is that it helps to ameliorate the judgment-proof problem as it applies to the injurer, as noted in the previous section.[37] Under vicarious liability, the vicariously liable party's assets are at risk as well as the injurer's, giving the vicariously liable party a motive to reduce risk or to moderate the injurer's level of activity. There are various ways in which the vicariously liable party can affect risk or the level of activity. (a) Vicariously liable parties may be able to affect the behavior of injurers in some direct manner, such as where a parent prevents a child from using a speedboat, where an employer does not allow an employee to transport hazardous wastes until the employee has been through a training course in handling the wastes, or where a contractor requires a subcontractor to use a particular safety device. Here, note, the control by the vicariously liable party may be over the level of care (as in the case of the employee who must take a training course) or over the level of activity (as in the case of the child who is prevented from using the speedboat). (b) Another possibility is that vicariously liable parties can themselves take precautions that alter the risk that injurers present. For example, the employer can purchase a tanker truck to transport hazardous wastes that is safer than some other kind of tanker truck (perhaps it has a double-walled container). (c) Sometimes vicariously liable parties may control participation in activities because they act as "gatekeepers"; they are able to prevent injurers from engaging in their activity by withholding financing or a required service. Thus, a vicariously liable lender or parent corporation may withhold funding to an injurer; or a vicariously liable architect may not agree to sign a statement that a hotel is safe from collapse in an earthquake and thus effectively deny the hotel from doing business if such a statement is legally required.

The main disadvantage of vicarious liability is that it increases litigation costs, since it means that vicariously liable parties can be sued as well as the injurer. The increase in litigation cost must be weighed against the improvement in incentives. Hence, imposing vicarious liability on a homeowner for an accident caused by a delivery truck that is bringing something to his home would probably not be worthwhile, since the extra incentive gained would be modest. Another disadvantage is that it leads potential contracting parties, who might become vicariously liable were they to make a contract, to expend effort determining the degree of their liability exposure.[38]

[37] Both reasons often apply jointly; for instance, children and employees are often judgment proof as well as in possession of less information than vicariously liable parties. Other reasons for vicarious liability are sometimes emphasized. For example, where a firm is held vicariously liable for its employees, a victim need not have to prove which employee actually caused the harm suffered.

[38] Another disadvantage, of a subtle nature, is that a vicariously liable contracting party will charge a party with low assets an amount reflecting the vicariously liable party's liability exposure; this charge will itself reduce the assets of the low asset party, exacerbating the judgment proof problem in regard to him; see Pitchford (1995).

Literature on vicarious liability originates with Kornhauser (1982) and Sykes (1981, 1984).[39]

12. Nonpecuniary harm

It is often the case that the victim of an accident suffers a nonpecuniary harm, such as when an irreplaceable photo album is destroyed in a fire, a person suffers a disabling injury, or a scenic beach is ruined by an oil spill. In these cases, the harm is called nonpecuniary because the loss is not replaceable with a good or service that can be purchased on a market. Let us sketch the implications of such nonpecuniary harms for the analysis of the liability system.

12.1. Incentives in the risk-neutral case

Suppose initially that individuals are risk neutral. In this case, it is natural to assume that the victim's loss is the sum of a pecuniary component h and a nonpecuniary component t. Thus, the optimal level of care x^* is the x minimizing $x+p(x)(h+t)$; the optimal level z^* at which to engage in the activity is the z maximizing $b(z)-z[x^*+p(x^*)(h+t)]$; and so forth. Hence, it is evident that the nonpecuniary harm should influence damages and the due care standard: the optimal level of damages is $d = h + t$; and due care should be x^* and not the lower level of care minimizing only $x + p(x)h$.

However, courts often face an informational problem in determining nonpecuniary harm t, and one that is more difficult than that experienced in determining pecuniary harm. Pecuniary harm is measurable since, by definition, it is equal to the cost of purchasing a replacement good. If a car is damaged beyond repair, the same car (or essentially the same one) can generally be purchased on the market, so the court needs only to estimate the cost of buying the replacement vehicle. In contrast, nonpecuniary harm is not usually associated with a market observable and, moreover, varies, perhaps greatly, across individuals. The disutility of losing an irreplaceable family photo album might be very different from one person to the next. Thus, courts are likely to err in computing damages for nonpecuniary harm (and in ascertaining the proper due care level when such harm is involved).

12.2. Incentives and insurance in the risk-averse case; fines as a supplement to liability

Now suppose that victims are risk averse and for simplicity that a victim's utility is of the form $u(y)$ plus a nonpecuniary utility component, where $u(y)$ is utility from wealth

[39] See also Boyd and Ingberman (1997), Boyer and Laffont (1997), Feess (1999), Hansmann and Kraakman (1991), Kraakman (1986, 1998), Pitchford (1995), and Shavell (1987, pp. 170–177, 182–185).

y and u is concave. If there is an accident causing a pecuniary harm of h and a utility loss of t, the victim's utility would be $u(y - h) - t$; whereas if there is no accident, his utility is $u(y)$. Before considering liability and incentives, observe that the optimal insurance policy for a victim facing a loss of h and t is a policy that covers only the pecuniary loss h. For, as is well known, optimal insurance coverage is such that the marginal utility of wealth is the same if an accident occurs as if it does not; and since the utility loss t does not affect the utility of wealth, purchasing insurance against it would only lower expected utility.[40] Another way to put this point is that although nonpecuniary losses may involve substantial disutility (consider the death of a child or a parent), they may not create a need for money (do not raise the marginal utility of money), so it makes no sense to insure against them.

This point about the optimal insurance purchases against nonpecuniary losses suggests that the first-best solution to the accident problem when losses involve nonpecuniary components and victims are risk averse has the following character. Victims receive an amount equal to their pecuniary losses alone; but injurer risk-reducing behavior reflects nonpecuniary as well as pecuniary harm.

The first-best solution cannot be achieved under strict liability for the simple reason that if victims receive optimal insurance coverage equal to pecuniary harm, incentives of injurers will be too low; and if injurers pay damages sufficient to give them proper incentives, victims' insurance coverage will be excessive. The second-best level of damages strikes an implicit compromise between the goal of providing proper incentives to injurers and optimal insurance for victims.

However, the first-best solution can be achieved under strict liability if damages are supplemented by fines collected by the state. For then damages can be set equal to the pecuniary harm h, optimally insuring victims, and the fine can be set high enough so as to provide, together with damages, optimal incentives to reduce risk for injurers. (One way to appreciate the superiority of the system of fines is that it can be shown to be Pareto superior to any liability system. In a fully-specified model, the fine revenue can benefit individuals, such as by reducing their taxes or by allowing a lump sum payment to be made to individuals ex ante.)

Under the negligence rule, if it is perfectly functioning, the first-best solution can be achieved without the use of fines. For if damages are set high enough to induce due care of x^*, victims will bear their losses and will purchase accident insurance covering only their pecuniary harm h. However, if there are errors in the negligence determination, then, as under strict liability, a system with fines as a supplement to liability will tend to be superior to the pure liability system.

The use of supplemental fines would tend to resolve certain tensions arising from application of standard principles of tort law in calculating damages. It is frequently observed, for example, that because damages for death tend to be based on the present value of foregone earnings, payments for the death of children (whose earnings are

[40] See Arrow (1974) and Cook and Graham (1977).

many years in the future) or for the elderly are low. Such damages lead to inadequate deterrence (and also may seem jarring to the public), even though the damages are not inappropriate from the standpoint of insurance (the death of a child or of an elderly person usually does not create the need for money). The use of fines could help to increase deterrence without altering what victims receive.[41]

Part C: Methods of controlling risk

13. Liability versus other methods of controlling risk

13.1. *Methods of control*

Liability for accidents is one among a number of methods that society may employ to control potentially harmful behavior. Another is direct regulation, under which the state restricts permissible behavior, for example, by requiring that factories install smoke arrestors. Closely related to state regulation is the legal injunction, whereby a potential victim can enlist the power of the state to force a potential injurer to take steps to prevent harm or to cease his activity; the injunction is effectively privately-initiated regulation. Society can also make use of the corrective tax, under which a party pays the state an amount equal to the expected harm he causes, a standard example being the expected harm due to the discharge of a pollutant into a lake.[42] In fact, liability and regulation are the preeminent tools that society utilizes to control risky activities; use of corrective taxes is unusual in a relative sense.

13.2. *Comparison*

I will now sketch several factors bearing on the relative desirability of these methods of controlling externalities. The review of factors will show that any of the methods (or a combination) could be the best, depending on the context.[43]

One factor of relevance is the quality of the state's information. If the state has complete information, that is, knows injurers' benefits from acts, costs of precautions and their effectiveness, and the harm that injurers would cause, then all of the methods of

[41] The subject of liability, risk-bearing, and non-pecuniary harms was first studied in Spence (1977), who pointed out the advantage of fines as a supplement to liability. See also Calfee and Rubin (1992), Shavell (1987, pp. 231–235, 247–254), and Viscusi and Evans (1990).

[42] Other tools include subsidies for taking precautions and fines for causing harm, but for simplicity only the four methods mentioned in the text are considered here.

[43] This section follows Shavell (1993). There are a number of articles that focus on comparisons between pairs of methods of controlling externalities. See Calabresi and Melamed (1972), Ellickson (1973), Kaplow and Shavell (1996a, 2002), Kolstad et al. (1990), Polinsky (1980a), Shavell (1984a, 1984b), Weitzman (1974), and Wittman (1977).

control allow achievement of optimality. If the state's information is imperfect, the state will not be able to calculate which actions (such as installing a smoke arrestor) are desirable and thus sometimes will err. Hence, methods of control that require the state to determine optimal actions—namely, regulation, the injunction, and liability based on the negligence rule—will not perform well. In contrast, as long as the state can ascertain expected harm, we know that it can induce injurers to act optimally under the corrective tax; and as long as the state can determine the actual harm, we know that it an induce optimal behavior under strict liability; for under either the corrective tax or strict liability, injurers will balance their costs of precautions and benefits from engaging in activities against the reduction in expected harm that would be brought about by taking precautions and modifying their levels of activity.

This basic informational argument favoring corrective taxes or strict liability over regulation, the injunction, and the negligence rule applies despite possible uncertainty about the magnitude of harm. The reason, essentially, is that under corrective taxes, the state only needs to estimate expected harm and set the tax equal to it, and under strict liability, the state only needs to allow harm to occur and set damages equal to actual harm, whereas under the other methods the state must also estimate precaution costs and/or injurers' benefits from engaging in their activities. The present point about the superiority of corrective taxes (and of strict liability) holds notwithstanding Weitzman (1974), which suggests that regulation may be superior to corrective taxation.[44] The state's information also bears on the relative appeal of the corrective taxes versus strict liability. Under the corrective tax, the expected harm must be known, such as the expected harm that would be caused by pollution or by leaving ice on sidewalks. Under strict liability, only the actual harm need be known. When, as is often the case, actual harm is more readily measured than expected harm (arguably the case for icy sidewalks), strict liability enjoys an informational advantage over the corrective tax. In some situations, though, expected harm may be easier to determine than actual harm (especially when harm is difficult to trace to its origin).

A second factor of relevance to the comparison of methods of control is the information available to victims. For many externalities, victims have better information than the state about who is causing harm or about its extent—because they are the parties who actually suffer the harm. Hence, one would suppose that victims are the most appropriate enforcement agents, suggesting the desirability of the liability tool or the injunction. For some externalities, however, victims may have poor information about harm, worse than the state's. Consider, for instance, that victims may be unaware that harm is even occurring when they are exposed to odorless and invisible carcinogenic agents. In such circumstances, the state would often be a better enforcer than individuals, and use of regulation or corrective taxes might be advantageous.

[44] Weitzman's (1974) conclusion that regulation could be superior to taxation rests on the assumption that the tax rate does not depend on the quantity of pollution. If, as would usually be easy to implement, the tax rate can vary with the quantity of pollution, taxes are superior to regulation; see Roberts and Spence (1976) and Kaplow and Shavell (2002).

A third factor concerns the level of activity of an injurer (how much a firm produces, how many miles a person drives), as opposed to the precautions an injurer takes given the level of activity (whether a firm uses a smoke scrubber while producing, whether a person exercises care when driving). Regulation and the negligence rule are most often concerned with precautions taken but not with the level of activity (a firm may be required by regulation to install smoke scrubbers or to take certain safety precautions but not to reduce its output). Thus injurers may not have incentives to moderate their level of activity although that would be desirable (their activity may result in harm despite the exercise of optimal precautions—even with smoke scrubbers, some pollution will result). In contrast, under the corrective tax and strict liability, injurers pay for harm done, so that they will optimally moderate their level of activity (as well as efficiently choose their level of precautions).

A fourth pertinent factor, noted above, is the ameliorative behavior of victims. Under regulation, the injunction, and corrective taxation, victims are not compensated for their harm, so that victims have a natural incentive to take optimal precautions or to alter their activity. Similarly, under the negligence rule, victims have an incentive to take precautions, since injurers will tend to be led to behave nonnegligently so will not compensate victims. Under strict liability, victims incentives are not as good; although the defense of contributory negligence gives victims an incentive to take precautions, they are not led to moderate their level of activity (or any aspect of behavior not included in the contributory negligence determination).

Still another factor is administrative costs, the costs borne by the state in applying a legal rule and the legal and related costs borne by the affected parties. Liability rules possess a general administrative cost advantage over regulation in that under liability rules, administrative costs are incurred only if harm is done. This advantage may be significant when the likelihood of harm is small. Nevertheless, administrative costs will sometimes be lower under other approaches. For example, compliance with a regulation may readily be detected in some circumstances (determining whether factory smokestacks are sufficiently high would be easy) and also may be accomplished through random monitoring, saving enforcement resources. Also, imposing corrective taxes might be inexpensive. Notably, suppose that they are levied at the time of the purchase of a product. In contrast, liability rules might be expensive to employ. For example, demonstrating the source of a particular harm and its extent may be difficult. Also, when industrial pollution affects millions of individuals on an ongoing basis, the cost of a continuous flow of individual suits (or even class actions) that measure damages victim-by-victim is likely to be in excess of the cost of alternatives.

Last, the ability of injurers to pay for harm is of relevance. For liability rules to induce potential injurers to behave appropriately, injurers must have assets sufficient to make the required payments; otherwise they will have inadequate incentives to reduce harm, as discussed in section 10. Where inability to pay is a problem, bonding requirements may be helpful, and regulation may become more appealing (although it may need to be enforced through the threat of nonmonetary, criminal sanctions). In addition, corrective taxes have an advantage over liability rules when harm is probabilistic because, under the corrective tax, an injurer would pay only the expected harm (with certainty) rather than the higher actual harm (with a probability); see section 10.4.

13.3. Resolution of externalities through bargaining by affected parties

Parties affected by unregulated externalities will sometimes have the opportunity to make mutually beneficial agreements with those who generate the externalities. In the classic example, if a firm's pollution causes harm of $1,000 that can be prevented by installing a smoke scrubber that costs $100, then, in the absence of any legal obligation on the firm, one might expect a potential victim of pollution to pay the firm an amount between $100 and $1,000 to install the scrubber.[45]

If it is posited that there are no obstacles to reaching a mutually beneficial agreement concerning externalities, then that will occur. This tautology is one version of the Coase Theorem; Coase (1960) stressed the point that externality problems could be remedied through private bargains. A closely related version of the Coase Theorem asserts that the outcome regarding the externality—whether a smoke scrubber is installed or instead pollution is generated—does not depend on the legal rule that applies. For example, if the scrubber costs $100 and there is no law that controls pollution, a bargain as we have described it will come about and the scrubber will be installed; and likewise if there is a law that leads to installation of the scrubber, the same will happen. The outcome, however, might be affected by the legal rule because of the level of wealth of parties. Most obviously, the potential victims might not have assets sufficient to pay for the scrubber, in which case the scrubber would not be installed unless a legal rule leads to this; moreover, legal rules may affect the distribution of wealth and thus the demand for goods, including that of being free from pollution.[46]

In any case, there are many obstacles to bargaining. Bargaining may fail to occur when victims are numerous and face collective action problems in coming together. This is often the situation with respect to victims of industrial pollution, for example. Similarly, in important contexts, bargaining will be impractical because victims will not know in advance who will injure them; this is the case for automobile accidents and most other accidents between strangers. Another reason that bargaining may not occur is that victims might not know that they are exposed to a risk. Also, the cost of bargaining between just one potential victim and one potential injurer who know of each other can discourage them from engaging in the process. If these reasons do not apply and victims and injurers do engage in bargaining, asymmetry of information may lead to bargaining impasses. Altogether, these problems that reduce the likelihood of bargaining occurring, and also its success if it does take place, make the importance of legal rules to remedy externalities substantial.

[45] For experimental studies on bargaining and entitlements, see, for example, Hoffman and Spitzer (1982) and Croson and Johnston (2000).

[46] The outcome following from a legal rule might also be affected by an "endowment effect," under which individuals' valuations depend on whether or not they originally enjoy legal protection. See, for example, Kahneman et al. (1990).

14. Conclusion

The subject of liability for accidents as a device for controlling participation in risky activities and for inducing the taking of precautions when parties engage in them has been reviewed here. Economic analysis of this subject has been undertaken for only a relatively short time, and many avenues for theoretical research are apparent, concerning, for example, the relationship between liability insurance and the liability system, the ramifications of the judgment-proof problem, and the effect of liability on investigation of risks. More generally, as the discussion in section 13 should have indicated, analysis of the comparative strengths of (and of complementarities between) liability and other devices for controlling accidents seems to be of particular importance.

References

Arrow, K.J. (1974). “Optimal insurance and generalized deductibles”. Scandinavian Actuarial Journal 1974, 1–42.

Bar-Gill, O., Ben-Shahar, O. (2003). “The uneasy case for comparative negligence”. American Law and Economics Review 5, 433–469.

Boyd, J., Ingberman, D.E. (1997). “The search for deep pockets: is “extended liability expensive liability?” Journal of Law, Economics, and Organization 13, 232–258.

Boyer, M., Laffont, J.-J. (1997). “Environmental risks and bank liability”. European Economic Review 41, 1427–1459.

Brown, J.P. (1973). “Toward an economic theory of liability”. Journal of Legal Studies 2, 323–349.

Calabresi, G. (1970). The Costs of Accidents. Yale University Press, New Haven.

Calabresi, G. (1975). “Concerning cause and the law of torts”. University of Chicago Law Review 43, 69–108.

Calabresi, G., Melamed, A.D. (1972). “Property rules, liability rules, and inalienability: one view of the cathedral”. Harvard Law Review 85, 1089–1128.

Calfee, J.E., Rubin, P.H. (1992). “Some implications of damage payments for nonpecuniary losses”. Journal of Legal Studies 21, 371–411.

Coase, R.H. (1960). “The problem of social cost”. Journal of Law and Economics 3, 1–44.

Cook, P., Graham, D. (1977). “The demand for insurance and protection: the case of irreplaceable commodities”. Quarterly Journal of Economics 91, 143–156.

Cooter, R.D. (1989). “Punitive damages for deterrence: when and how much?” Alabama Law Review 40, 1143–1196.

Cooter, R.D., Ulen , T.S. (1986). “An economic case for comparative negligence”. New York University Law Review 61, 1067–1110.

Craswell, R., Calfee, J.E. (1986). “Deterrence and uncertain legal standards”. Journal of Law, Economics, & Organization 2, 279–303.

Croson, R., Johnston, J.S. (2000). “Experimental results on bargaining under alternative property rights regimes”. Journal of Law, Economics, & Organization 16, 50–73.

Cummins, J.D., Phillips, R.D., Weiss, M.D. (2001). “The incentive effects of no-fault automobile insurance”. Journal of Law and Economics 44, 427–464.

Danzon, P.M. (1985). Medical Malpractice: Theory, Evidence, and Public Policy. Harvard University Press, Cambridge, MA.

Daughety, A.F., Reinganum, J.F. (1997). “Everybody out of the pool: products liability, punitive damages, and competition”. Journal of Law, Economics, & Organization 13, 410–432.

Dewees, D., Duff, D., Trebilcock, M. (1996). Exploring the Domain of Accident Law: Taking the Facts Seriously. Oxford University Press, New York.

Diamond, P.A. (1974). “Single activity accidents”. Journal of Legal Studies 3, 107–164.

Diamond, P.A. (2002). “Integrating punishment and efficiency concerns in punitive damages for reckless disregard of risks to others”. Journal of Law, Economics, & Organization 18, 117–139.

Dobbs, Dan, B. (2000). The Law of Torts. West, St. Paul.

Edlin, A. (1994). “Efficient standards of due care: should courts find more parties negligent under comparative negligence?”. International Review of Law and Economics 14, 21–34.

Eisenberg, T., Goerdt, J., Ostrom, B.J., Rottman, D., Wells, M.T. (1997). “The predictability of punitive damages”. Journal of Legal Studies 26, 623–661.

Ellickson, R.C. (1973). “Alternatives to zoning: covenants, nuisance rules, and fines as land use controls”. University of Chicago Law Review 40, 681–781.

Emons, W. (1990). “Efficient liability rules for an economy with non-identical individuals”. Journal of Public Economics 42, 89–104.

Emons, W., Sobel, J. (1991). “On the effectiveness of liability rules when agents are not identical”. Review of Economic Studies 58, 375–390.

Farber, H.S., White, M.J. (1991). “Medical malpractice: an empirical examination of the litigation process”. Rand Journal of Economics 22, 199–217.

Feess, E. (1999). “Lender liability for environmental harm: an argument against negligence based rules”. European Journal of Law and Economics 8, 231–250.

Goldberg, V.P. (1974). “The economics of product safety and imperfect information”. Bell Journal of Economics 5, 683–688.

Grady, M.F. (1983). “A new positive economic theory of negligence”. Yale Law Journal 92, 799–829.

Green, J. (1976). “On the optimal structure of liability laws”. Bell Journal of Economics 7, 553–574.

Grossman, S.J. (1981). “The informational role of warranties and private disclosure about product quality”. Journal of Law and Economics 24, 461–483.

Hamada, K. (1976). “Liability rules and income distribution in product liability”. American Economic Review 66, 228–234.

Hansmann, H., Kraakman (1991), R. (1991). “The uneasy case for limiting shareholder liability in tort”. Yale Law Journal 100, 1879–1934.

Hoffman, E., Spitzer, M. (1982). “The Coase theorem: some experimental tests”. Journal of Law and Economics 25, 73–98.

Huberman, G., Mayers, D., Smith, C. (1983). “Optimal insurance policy indemnity schedules”. Bell Journal of Economics 14, 415–426.

Hylton, K.N. (1990). “The influence of litigation costs on deterrence under strict liability and under negligence”. International Review of Law and Economics 10, 161–171.

Jerry, R.H. (1996). Understanding Insurance Law. Matthew Bender, New York.

Jost, P.-J. (1996). “Limited liability and the requirement to purchase insurance”. International Review of Law and Economics 16, 259–276.

Kahan, M. (1989). “Causation and incentives to take care under the negligence rule”. Journal of Legal Studies 18, 427–447.

Kahneman, D., Knetsch, J.L., Thaler, R.H. (1990). “Experimental tests of the endowment effect and the Coase theorem”. Journal of Political Economy 98, 1325–1348.

Kakalik, J.S., Pace, N.M. (1986), Costs and Compensation Paid in Tort Litigation, Report R-3391-ICJ. Rand Corporation, Santa Monica, CA.

Kaplow, L. (1986). “Private versus social costs in bringing suit”. Journal of Legal Studies 15, 371–385.

Kaplow, L., Shavell, S. (1996a). “Property rules versus liability rules: an economic analysis”. Harvard Law Review 109, 713–790.

Kaplow, L., Shavell, S. (1996b). “Accuracy in the assessment of damages”. Journal of Law and Economics 39, 191–210.

Kaplow, L., Shavell, S. (2001). “Any non-welfarist method of policy assessment violates the Pareto principle”. Journal of Political Economy 109, 281–286.

Kaplow, L., Shavell, S. (2002). “On the superiority of corrective taxes to quantity regulation”. American Law and Economics Review 4, 1–17.

Karpoff, J.M., Lott, J.R. (1999). “On the determinants and importance of punitive damage awards”. Journal of Law and Economics 42, 527–573.

Keeton, W.R., Kwerel, E. (1984). “Externalities in automobile insurance and the underinsured driver problem”. Journal of Law and Economics 27, 149–179.

Kessler, D., Rubinfeld, D.L. (2007). “Empirical Study of the Common Law and Legal Process”. In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics, vol. 1. North-Holland, Amsterdam.

Kolstad, C.D., Ulen, T.S., Johnson, G.V. (1990). “Ex post liability for harm vs. ex ante safety regulation: substitutes or complements?” American Economic Review 80, 888–901.

Kornhauser, L.A. (1982). “An economic analysis of the choice between enterprise and personal liability for accidents”. California Law Review 70, 1345–1392.

Kraakman, R. (1986). “Gatekeepers: the anatomy of a third-party enforcement strategy”. Journal of Law, Economics, & Organization 2, 53–104.

Kraakman, R. (1998). “Limited shareholder liability”. In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 3. Macmillan, London, pp. 648–654.

Landes, E. (1982). “Insurance, liability, and accidents: a theoretical and empirical investigation of the effect of no-fault accidents”. Journal of Law and Economics 25, 49–65.

Landes, W.M., Posner, R.A. (1980). “Joint and multiple torts: an economic analysis”. Journal of Legal Studies 9, 517–556.

Landes, W.M., Posner, R.A. (1987). The Economic Structure of Tort Law. Harvard University Press, Cambridge, MA.

Menell, P.S. (1983). “A note on private versus social incentives to sue in a costly legal system”. Journal of Legal Studies 12, 41–52.

Oi, W. (1973). “The economics of product safety”. Bell Journal of Economics 4, 3–28.

Ordover, J.A. (1978). “Costly litigation in the model of single activity accidents”. Journal of Legal Studies 7, 243–261.

Pitchford, R. (1995). “How liable should a lender be? The case of judgment-proof firms and environmental risk”. American Economic Review 85, 1171–1186.

Polborn, M.K. (1998). “Mandatory insurance and the judgment-proof problem”. International Review of Law and Economics 18, 141–146.

Polinsky, A.M. (1980a). “Resolving nuisance disputes: the simple economics of injunctive and damage remedies”. Stanford Law Review 32, 1075–1112.

Polinsky, A.M. (1980b). “Strict liability vs. negligence in a market setting”. American Economic Review (Papers and Proceedings) 70, 363–367.

Polinsky, A.M. (1997). “Are punitive damages really insignificant, predictable, and rational?” Journal of Legal Studies 26, 663–677.

Polinsky, A.M., Che, Y.-K. (1991). “Decoupling liability: optimal incentives for care and litigation”. Rand Journal of Economics 22, 562–570.

Polinsky, A.M., Rubinfeld, D.L. (1988). “The welfare implications of costly litigation for the level of liability”. Journal of Legal Studies 17, 151–164.

Polinsky, A.M., Shavell, S. (1998). “Punitive damages: an economic analysis”. Harvard Law Review 111, 869–962.

Posner, R.A. (1972). Economic Analysis of Law, 1st edn. Little, Brown and Company, Boston.

Priest, G.L. (1981). “A theory of the consumer product warranty”. Yale Law Journal 90, 1297–1352.

Roberts, M.J., Spence, M. (1976). “Effluent charges and licenses under uncertainty”. Journal of Public Economics 5, 193–208.

Rose-Ackerman, S., Geistfeld, M. (1987). “The divergence between the social and private incentives to sue: a comment on Shavell, Menell, and Kaplow”. Journal of Legal Studies 16, 483–491.

Rosenberg, D. (1984). “The causal connection in mass exposure cases: a ‘public law’ vision of the tort system”. Harvard Law Review 97, 849–929.

Rubinfeld, D.L. (1987). “The efficiency of comparative negligence”. Journal of Legal Studies 16, 375–394.

Sarath, B. (1991). “Uncertain litigation and liability insurance”. Rand Journal of Economics 22, 218–231.

Schwartz, A., Wilde, L.L. (1979). "Intervening in markets on the basis of imperfect information: a legal and economic analysis". University of Pennsylvania Law Review 127, 630–682.

Shavell, S. (1980a). "An analysis of causation and the scope of liability in the law of torts". Journal of Legal Studies 9, 463–516.

Shavell, S. (1980b). "Strict liability versus negligence". Journal of Legal Studies 9, 1–25.

Shavell, S. (1982a). "On liability and insurance". Bell Journal of Economics 13, 120–132.

Shavell, S. (1982b). "The social versus the private incentive to bring suit in a costly legal system". Journal of Legal Studies 11, 333–339.

Shavell, S. (1984a). "A model of the optimal use of liability and safety regulation". Rand Journal of Economics 15, 271–280.

Shavell, S. (1984b). "Liability for harm versus regulation of safety". Journal of Legal Studies 13, 357–374.

Shavell, S. (1985). "Uncertainty over causation and the determination of civil liability". Journal of Law and Economics 28, 587–609.

Shavell, S. (1986). "The judgment proof problem". International Review of Law and Economics 6, 45–58.

Shavell, S. (1987). Economic Analysis of Accident Law. Harvard University Press, Cambridge, MA.

Shavell, S. (1993). "The optimal structure of law enforcement". Journal of Law and Economics 36, 255–287.

Shavell, S. (1997). "The fundamental divergence between the private and the social motive to use the legal system". Journal of Legal Studies 26, 575–612.

Shavell, S. (1999). "The level of litigation: private versus social optimality". International Review of Law and Economics 19, 99–115.

Shavell, S. (2000). "On the social function and the regulation of liability insurance". Geneva Papers on Risk and Insurance, Issues and Practice 25, 166–179.

Shavell, S. (2005). "Minimum asset requirements and compulsory liability insurance as solutions to the judgment-proof problem". Rand Journal of Economics 36, 63–77.

Sloan, F.A. (1998). "Automobile accidents, insurance, and tort liability". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 1. Macmillan, London, pp. 140–144.

Spence, M. (1977). "Consumer misperceptions, product failure, and product liability". Review of Economic Studies 44, 561–572.

Spier, K.E. (1994). "Settlement bargaining and the design of damage awards". Journal of Law, Economics, & Organization 10, 84–95.

Spier, K.E. (1997). "A note on the divergence between the private and the social motive to settle under a negligence rule". Journal of Legal Studies 26, 613–621.

Spier, K.E. (2007). "Litigation". In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics, vol. 1. North-Holland, Amsterdam.

Summers, J. (1983). "The case of the disappearing defendant: an economic analysis". University of Pennsylvania Law Review 132, 145–185.

Sykes, A.O. (1981). "An efficiency analysis of vicarious liability under the law of agency". Yale Law Journal 91, 168–206.

Sykes, A.O. (1984). "The economics of vicarious liability". Yale Law Journal 93, 1231–1282.

Tillinghast-Towers Perrin (2000). U.S. Tort Cases, 2000: Trends and Findings on the Costs of the U.S. Tort System. Tillinghast-Towers Perrin, Chicago.

Tunc, A. (1974). "Introduction". In: Torts. International Encyclopedia of Comparative Law, vol. II. J.C.B. Mohr, Tübingen, Germany, Chapter I.

Viscusi, W.K., Evans, W. (1990). "Utility functions that depend on health status: estimates and economic implications". American Economic Review 80, 353–374.

Weitzman, M.L. (1974). "Prices vs. quantities". Review of Economic Studies 41, 477–491.

Wittman, D. (1977). "Prior regulation versus post liability: The choice between input and output monitoring". Journal of Legal Studies 6, 193–212.

Chapter 3

PROPERTY LAW

DEAN LUECK

Department of Agricultural & Resource Economics and College of Law, University of Arizona

THOMAS J. MICELI

Department of Economics, University of Connecticut

Contents

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)01003-1

Abstract

This chapter examines the economics of property rights and property law. It shows how the economics of property rights can be used to understand fundamental features of property law and related extra-legal institutions. The chapter examines both the rationale for legal doctrine, and the effects of legal doctrine regarding the exercise, enforcement, and transfer of rights. It also examines various property rights regimes including open access, private ownership, common property, and state property. The guiding questions are: How are property rights established? What explains the variation in the types of property rights? What governs the use and transfer of rights? And, how are property rights enforced? In answering these questions we argue that property rights and property law can be best understood as a system of societal rules designed to maximize social wealth. They do this by creating incentives for people to maintain and invest in assets, which leads to specialization and trade.

Keywords

Coase theorem, property rights, property law, transaction costs, externality

JEL classification: D23, D62, K11, K23

1. Introduction

This chapter examines the economics of property rights and property law. The purpose is to show how the economics of property rights can be used to understand fundamental features of property law and related extra-legal institutions. The chapter will examine both the rationale for, and the effects of, legal doctrine. The guiding questions are: How are property rights established? What explains the variation in the types of property rights? What governs the use and transfer of rights? How are property rights enforced? In answering these questions we argue that property rights and property law can be best understood as a system of societal rules designed to maximize social wealth. They do this by creating incentives for people to maintain and invest in assets, which leads to specialization and trade.

1.1. Property rights and property law

Property rights have been a subject of discussion among philosophers as well as political and legal scholars long before economists began to examine their origins and consequences. Property rights were discussed and implicitly understood by ancient Greek and Roman writers, but it is perhaps Hobbes (1651) who first discusses property in a manner recognizable to modern economists. Indeed Hobbes' 'state of nature' can be viewed as open access dissipation. A wide range of Enlightenment thinkers such as Blackstone (1766), Hume (1739–1740), Locke (1690), and Smith (1776) also discussed property rights. Though they varied in their treatments all considered property rights as fundamental social institutions for creating wealth and preventing strife.

We define property rights as the ability (or expected ability) of an economic agent to use an asset (Allen, 1999; Barzel, 1997; Shavell, 2004). As Demsetz (1967) notes in one of the classic early economic analyses, property rights represent a social institution that creates incentives to efficiently use assets, and to maintain and invest in assets. They may or may not be enforced by courts and because the actions of courts are costly, legal rights are but a subset of economic property rights. In addition to law (and statutorily-based regulations enforced by administrative agencies), property rights may be enforced by custom and norms (e.g., Ellickson, 1991), and by markets through repeated transactions.[1]

Property law is the body of court enforced rules governing the establishment, use, and transfer of rights to land and those assets attached to it such as air, minerals, water, and wildlife.[2] Intellectual property law similarly details the conditions under which

[1] Enforcement may also be by violence as with the Mafia (Gambetta, 1993).

[2] Merrill and Smith (2001) and Arruñada (2003), however, argue that the *in rem* nature of property ('real rights' versus *in personam* 'use rights' typical in contracts) is an important distinction that property rights economists have overlooked since Coase (1960). Merrill and Smith (2001, p. 375) argue that '[t]he simplifying assumptions [of economists] introduce blind spots that can limit the ability of law and economics scholars to explain the institution of property.' They cite doctrines like the *numerus clausus* principle, and Arruñada notes the important of title systems as *in rem* enforcement institutions.

the courts enforce rights to intellectual assets. In this framework, virtually all, if not all, branches of law are 'property rights law.' Labor law defines the court's role in enforcing rights to one's labor, contract law defines the rights of contracting parties, and so on. Because the economics of property rights originated with a focus on rights to land and associated natural resources (e.g. fisheries, pastures, water) the link between 'property law' and 'property rights' is firmly established. This chapter will develop this link by examining property rights generally and property law in particular. Yet, much of the analysis in this chapter is applicable to topics elsewhere in the handbook, though in many cases (e.g., contracts, torts) the literature has become so specialized that the connection to the economics of property rights might seem faint.

The economic analysis of property law is substantially less well developed than the economic analysis of contract law or tort law (for example, there is no generally applicable model), and this chapter reflects this state of the discipline. The economics of property rights, however, is well developed but mostly without a focus on property law.[3] The disconnection between the economics of property rights and the economics of property law is longstanding. For example, Demsetz's (1998) recent entry 'Property Rights' in the *The New Palgrave Dictionary of Economics and the Law* makes absolutely no mention of property law, and much of the economics of property rights literature remains ignorant of property law.[4] Similarly, property law scholarship often is ignorant of economics. This is not to say there has not been important work in property law with strong economic underpinnings (e.g., Ellickson, 1993; Epstein, 1985a; Heller, 1998; Merrill, 1986; Rose, 1990), but it is clear that economics has not yet penetrated property law as it has penetrated contract and tort law. While it is common for courses in contract law and tort law to be taught using economics as the guiding framework, an economics-based course in property law is almost unheard of. [5] In part, this chapter seeks to break down this division by bringing the two literatures together.

1.2. Property rights, transaction costs, and the Coase Theorem

The economics of property law begins with Coase (1960), who provides a property rights perspective on the problem of externalities, or 'social cost.'[6] Prior to Coase, economists viewed externalities as a source of market failure requiring government intervention to force the responsible party to curtail the harmful activity. Consider Coase's famous example of the rancher and farmer with adjacent plots of land. The rancher's cattle stray onto the farmer's land causing crop damage. If the rancher's profit, $\pi(h)$, and the amount of crop damage, $d(h)$, are functions of the rancher's herd size, h, then the first-best optimal herd size, h^*, maximizes $\pi(h) - d(h)$. That is, h^* solves

[3] The main exception to this is a deep theoretical literature on takings which is examined in section 8.

[4] Merrill and Smith (2001) note this also.

[5] An exception is Dwyer and Menell (1998).

[6] Another important, though little known early property rights contribution is that of Alchian (1965).

$\pi'(h) = d'(h)$.[7] This is also the choice that would be made by a single party acting as both the farmer and rancher, Coase's 'sole owner' solution. First-best then is synonymous with the zero transaction cost outcome. With separate parties, however, and the absence of a contract between the farmer and the rancher or some type of government intervention (a tax, fine, or regulation), the rancher would choose the herd size to maximize $\pi(h)$. This results in too many cattle because the rancher adds cattle until $\pi'(h) = 0$, which implies $h^r > h^*$. Thus, the rancher must pay a tax (or face liability) for the damage from straying cattle, or he will expand his herd beyond the efficient (first-best) size.

Note that this solution to the externality problem embodies a particular assignment of property rights—namely, that the farmer has the right to be free from crop damage. Another way to say this is that the farmer is labeled as the 'cause' of the harm and therefore must face liability. And if the property right (or the legal liability rule) is structured properly, the rancher will purchase the right to impose crop damage up to the point where the marginal profit from the last steer just equals the marginal damage, yielding an efficient herd size.

Coase's critique of this conventional, or 'Pigovian,' perspective on externalities is not that it is wrong per se, but that it is incomplete.[8] To illustrate, suppose that the rancher initially has the economic (and legal) right to impose crop damage without penalty. According to the Pigovian view, this would result in an excessive herd size because the rancher would expand the herd to h^r. But note that the farmer would be willing to pay up to $d'(h)$, his marginal damage, for each steer that the farmer removes from the herd in order to avoid crop damage, while the rancher would accept any amount greater than his marginal profit, $\pi'(h)$. Thus, if transaction costs are zero, the parties will contract to reduce the herd to the efficient size. In other words, the farmer will purchase the rights to the straying cattle, the reverse of what happened under the Pigovian solution. The outcome in both cases is therefore first-best. This conclusion has become known as the Coase Theorem[9], which can be stated in general terms as follows: When transaction costs are zero, the allocation of resources will be efficient regardless of the initial assignment of property rights.[10]

Coase challenged two assumptions implicit in the Pigovian view: first, that there is a unique cause of the harm, and second, that government intervention is necessary to internalize the externality. Coase noted that in general, and in the specific farmer-rancher

[7] We assume that $\pi'' < 0$ and $d'' > 0$, ensuring a unique optimum.

[8] This tradition is attributed to A.C. Pigou's *The Economics of Welfare* (1932).

[9] Stigler (1987) takes credit for calling this proposition the 'Coase Theorem.' Stigler also recounts a famous dinner at the home of Aaron Director where Coase convinced a formidable group of scholars (including Milton Friedman and Stigler) that his analysis was indeed correct.

[10] Typically economists have argued that the Coase Theorem is conditioned on the size of wealth effects. Barzel (1997), however, argues that wealth effects are likely to be trivial and not a condition of the Coase Theorem. He notes that the standard example of rights shifting without compensation itself violates the assumption of zero transaction costs. This is because zero transaction costs means that a rights shift would have to be accompanied by a payment to the original rights holder.

example, the cause of harm is 'reciprocal' in the sense that if either party is removed, the harm disappears.[11] Second, by noting that well defined ownership can lead to transactions, the range of 'solutions' extends to private contracting as well as government regulation and taxation.[12]

1.3. The impact of transaction costs: when does law matter?

Although there has been debate among economists and legal scholars on the significance of the Coase Theorem and its implications, Coase (1960, 1988) has been clear on this issue. Economic and legal institutions are important and have impacts because transaction costs are *not zero* and thus property rights are *not perfectly defined* (Allen, 1999; Barzel, 1997).[13] The Coase Theorem thus stresses the role of transaction costs in shaping the institutions, including law, that determine the allocation of resources. Seen as the costs of defining and enforcing property rights, transaction costs include enforcement costs, measurement costs, and moral hazard costs (Allen, 1998).

But Coase's insight goes further. Not only does the law matter for efficiency, as Demsetz (1972) explicitly points out, but the law itself is an economic choice, also expected to be driven by economic forces.[14] Indeed, Coase's (1960) discussion of nuisance law suggests an economic logic to the law in its assignment of property rights among various parties to these disputes, but its relevance extends beyond the study of externalities. It is concerned with the larger question of how property rights are established, the types of property rights regimes that are allowed, and the rules that govern the use and transfer of property rights. In this sense, property law is a complement to markets. This is the real contribution of Coase, and it will emerge as a theme throughout this chapter in a wide range of property law settings.

1.4. Outline of chapter

The remainder of the chapter is organized as follows. Section 2 develops the basic economic models of property rights that guide the analysis throughout the chapter. Section 3 examines the origin of rights. Section 4 follows with an analysis of the changes in property rights, or what has become known as the evolution of rights. Section 5 then examines various forms of voluntary exchange, including markets, leases, and inheritance.

[11] In technical terms Coase points out that most interesting actions (Y) depend on the inputs of both parties (a, b); that is, $Y = f(a, b)$. This 'bilateral' externality is discussed in section 7.1.

[12] This 'contractual' approach is discussed in section 4.4.

[13] Many scholars have called a case of zero transaction costs a 'Coasian world' but Coase (1988, p. 174) claims 'The world of zero transaction costs has often been described as a Coasian world. Nothing could be further from the truth. It is the world of modern economic theory, one which I was hoping to persuade economists to leave.'

[14] In another path breaking article, Coase (1937) uses a similar transaction cost argument to explain the boundary between markets and firms. Barzel and Kochin (1992) note the link between Coase's property rights and transaction costs theories.

Section 6 examines involuntary transfers of title by adverse possession and theft. Section 7 examines various means of internalizing externalities. Section 8 considers issues related to state (collective) ownership, as opposed to private ownership, of property, including the optimal scale of ownership and takings. Section 9 considers restrictions on the alienability of property. Finally section 10 concludes. Each section is a mix of formal and informal theory and application to law and related institutions. Throughout we try to make clear that the goal of the chapter is to use economics to illuminate the rationale for and effects of property law doctrine. Where possible we summarize the empirical literature or explain empirical applications. The sections are not symmetric, simply because the literature is not symmetric.[15]

2. Basic property rights models

Before examining property law doctrine it is appropriate to first examine the predominant types of property rights regimes and their economic structure. In this section we both describe the various types of property and examine the implications of these regimes for the use of and investment in assets. The models used in this section are fundamental to the later sections of the chapter that examine specific issues and legal doctrine.

As we noted above, there first systematic analysis of property rights began with the Enlightenment writers (e.g., Blackstone, Hume, Locke, Smith), but the formal modeling of the economics of property rights began with Frank Knight's (1924) analysis of public and private roads. Knight showed that a public road with no charge for access would be overused compared to the private road because users would not face the full cost of their actions. Gordon (1954) further developed Knight's preliminary model—establishing the now famous 'average product rule' for input use—in the context of an open ocean fishery where no one could be excluded.[16] Gordon's model was completed with Cheung's (1970) paper, which fully characterized the Nash equilibrium for an open access resource.

Our analysis of various property rights regimes will use a common set of notation in which a fixed asset (e.g., plot of land) is used in conjunction with a variable input (x) in order to produce a market output ($Y = f(x)$). If the input is available at a market wage of w, then the first-best use of the input ($x^*(w)$) must maximize $R = f(x) - wx$ and satisfy the first-order necessary condition $f'(x) = w$. The first-best value of the land is thus $V^* = \int_0^\infty R^*(x^*, t)e^{-rt}dt$, where r is the discount rate.[17] We start with open access, or a complete lack of property rights, and then, in turn, examine private property rights, common property, and mixed property rights regimes.[18]

[15] Because of its focus on specific assets, we make little use of the 'property rights theory of the firm' (e.g., Hart and Moore, 1990) which has become important in the economics of organization.

[16] Scott (1955) similarly shows the dissipation under open access and the private property solution.

[17] Each period's rent can be viewed as a steady state outcome.

[18] In law, since Roman times, open access resources have been called *res nullius*, or things unowned.

2.1. Open access

Assume there are n individuals who have unrestricted access to a resource such as a piece of land, and that output from the land (e.g., beef from grazing animals) is given by $Y = f(\sum_{i=1}^{n} x_i)$ where x_i is the effort of the i^{th} individual, $f'(\cdot) > 0$ and $f''(\cdot) < 0$, and the opportunity cost of effort is the market wage, w_i.[19] Each person's objective is to maximize his own rent subject to the constraint of open access, which means that each user can only capture (and own) the output in proportion to his share of effort.[20] This means each person must solve the following constrained maximization problem:

$$\begin{aligned}&\max_{x_i} R_i = f^i(x_i) - w_i x_i\\&\text{subject to } f^i = \left[x_i \Big/ \sum_{i=1}^{n} x_i\right] f\left(\sum_{i=1}^{n} x_i\right).\end{aligned} \tag{2.1}$$

Assuming that all users are homogeneous[21] ($w_i = w_j$, for all $i \neq j$), the Nash open access equilibrium is $x = x^{oa}(n, w_1, \ldots, w_n)$, which must satisfy the first-order necessary condition

$$(n - 1/n)\left(f\left(\sum_{i=1}^{n} x_i\right) \Big/ \sum_{i=1}^{n} x_i\right) = (1/n) f'\left(\sum_{i=1}^{n} x_i\right) = w_i, \quad i = 1, \ldots, n. \tag{2.2}$$

Equation (2.2), as Cheung shows, is indeed identical to Gordon's asserted average product equilibrium, but only in the limiting case of an infinite number of users with unrestricted access.[22] Thus, in the limit as $n \to \infty$, (2.2) becomes

$$\left(f\left(\sum_{i=1}^{n} x_i\right) \Big/ \sum_{i=1}^{n} x_i\right) = w, \tag{2.3}$$

which states that the open access equilibrium level of effort occurs where the average product equals the wage. More importantly, this limiting case also implies that rents are completely dissipated; or that, $\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} [f^i(x^{oa}) - wx^{oa}] = 0$. Similarly, the present value of the asset is also zero; that is, $V^{oa} = \int_0^{\infty} R(x^{oa}, t)e^{-rt} dt = 0$.

In this framework, the absence of property rights leads to overuse of the asset and complete dissipation of its value.[23] Complete dissipation is a limiting result, however,

[19] This production function captures the effect of competing users of the open access asset and is standard in the literature. Also, note that while ownership of the land is absent each person is assumed to have perfect ownership of themselves, their labor, and the product derived from the open access asset.

[20] This is a standard assumption but might be modified to explicitly distinguish use effort from violence effort.

[21] This has been the starting point with Knight, Gordon and Cheung.

[22] Equation (2.2) is actually a weighted average of average and marginal products. Brooks et al. (1999) show that Cheung's (1970) equilibrium holds in a dynamic setting.

[23] Hardin's (1968) famously named "tragedy of the commons" is a popularized version of this literature.

of the assumption of homogeneous users. If users are heterogeneous, dissipation under open access will be incomplete, and infra-marginal (low cost) users will earn rents (Libecap, 1989). The presence of rent under open access may be an important factor in preventing the establishment of rights to the open access resource because those earning rents will have incentives to maintain the open access regime.

2.2. *Private property rights*

Private ownership, as Knight first noted, is the straightforward solution to the open access problem.[24] Under the conditions of the Coase Theorem, the owner faces the full value and opportunity cost of asset use, he chooses the first-best level of use ($x^* < x^{oa}$), and generates $V^* > V^{oa} = 0$. The Coase Theorem also implies that, as long as property rights are well defined the organization of the asset's use will not matter: the owner may use the land himself, he may hire inputs owned by others, input owners may hire (or rent) the asset, or there may be a sharing arrangement between the asset owner and the input owners. In fact, under the conditions of zero transaction costs, any property regime (e.g., common property, state ownership) would generate the first-best use of the asset.

Not only does private ownership create incentives for optimal resource use, it also creates incentives for optimal asset maintenance and investment. With open access, no user has any incentive to use inputs that have a future payoff.[25] To see the effect on investment, consider a slightly modified version of the model above.[26] Let future output be $Y_{t+1} = f(x_t)$, where x_t is current investment, available at a market wage of w, and the interest rate is r. The first-best use of the input (x_t^*) must maximize $R = f(x_t)/(1 + r) - w_t x_t$ and satisfy the first-order necessary condition $f'(x_t)/(1 + r) = w_t$. This outcome is generated under perfect private ownership. Now let π be the probability of expropriation (because of imperfect property rights) of the future output, so that $(1 - \pi)$ is the probability the investor's output remains intact. The solution to the investment problem (x_t^π) is now to maximize $R = f(x_t)[(1 - \pi)/(1 + r)] - w_t x_t$ which must satisfy $f'(x_t)[(1 - \pi)/(1 + r)] = w_t$. This clearly implies less investment ($x_t^\pi < x_t^*$). Pure open access means that no investor could claim future output ($\pi = 1$), so $x_t^{oa} = 0$, and the rent from investment also equals zero.

In a recent article, Heller (1998) identifies a situation in which a large number of uncoordinated individuals have the right to exclude users, thus creating a regime in which assets are under-used because each rights holder can exercise a 'veto' over use.

[24] This was well understood by Hobbs, Bentham, Locke and Blackstone long ago.

[25] Writing before Adam Smith, Wm. Blackstone (Book II, Chapter 1, 1765) recognized this and wrote: 'And the art of agriculture, by a regular connexion and consequence, introduced and established the idea of a more permanent property in the soil, than had hitherto been received and adopted. It was clear that the earth would not produce her 'fruits in sufficient quantities, without the assistance of tillage: but who would be at the pains of tilling it, if another might watch an opportunity to seize upon and enjoy the product of his industry, art, and labor?'

[26] This is based on the detailed analysis of Bohn and Deacon (2000).

Because of the incentive to under use rather than over use the asset, Heller labeled this the 'anti-commons' and argues that many of the development problems in post-communist Europe are plagued with this problem of 'too many owners.' Buchanan and Yoon (2000) formalize Heller's idea and give it additional application in cases where competing bureaucracies can stifle development by exercising veto rights.[27] De Soto's (2000) documentation of the difficulties of operating in an economy heavily laden with overlapping bureaucracies is a similar application (as discussed in section 5.1.2). In yet another application, Anderson and Lueck (1992) study 'fractionated' ownership of land on American Indian reservations. They found that divided ownership of agricultural land (among large numbers of heirs to the original owner) led to dramatic reductions in the value of agricultural output.

It is not clear that the anti-commons phenomenon can usefully be regarded as a distinct property regime. The lack of investment incentives does not stem from 'too much ownership', but simply from severely divided interests in which unanimous agreement is required for decision.[28] In this sense, the anti-commons is like open access: too many people have access and thus no one can gain from optimal use. The difference seems to be that the decisions considered are investment decisions, so the anti-commons can be viewed as open access investment problem. The same land or apartments governed by 'too many' will be overused while investment will be suboptimal.

The empirical literature on private property rights is of two types.[29] First, there is a literature that attempts to measure the dissipation from open access and to compare resource use to that under private property. This rather small literature is dominated by studies on natural resources and especially of fisheries where open access regimes have been common (e.g., Agnello and Donnelley, 1975; Bottomley, 1963). These studies have estimated the deadweight losses from open access use and compared levels of asset use in open access regimes with those of private property and other limited access regimes. Second, there is a recent and growing literature on the effects of property rights security on resource use and investment. Much of this literature has focused on the investment effects of differences in legal title to land. In his survey article Besley (1998) notes that the econometric evidence for positive investment effects of more secure rights in developing countries is quite limited. These studies suffer, however, from data limitations (on both measures of investment and measures of property rights security) and from potential property rights endogeneity.[30] We expect more investment with better defined rights, but as we discuss in section 4, the choice of property rights regime can itself be influenced by investment levels or other correlated variables. Thus

[27] Parisi, Schulz, and Depoorter (2006) apply the idea of the anticommons to explain how the law limits fragmentation of rights.

[28] Heller and Eisenberg (1998) also call open access a problem of 'too many owners,' whereas the economic models examined above characterize this as a lack of ownership.

[29] There is also a property rights literature that focuses on differences between private and regulated firms (e.g., public utilities) that we do not discuss here.

[30] We return to this issue in the context of the choice of title system in section 5.1.1.

the econometric issue is how to find an instrument for property rights variables to isolate the effect of rights on investment. Still there is some compelling evidence for the importance of property rights in the cases of natural resource use (Bohn and Deacon, 2000), American Indian reservation agriculture (Anderson and Lueck, 1992), and urban residential land (Miceli et al., 2002).

2.3. *Common property rights*

In modern social science the term 'commons' or 'common property' originated in the analysis of what is now called open access. Yet, in law and custom common property has long meant, in stark contrast to open access, exclusive ownership by a group.[31] Common property regimes have been well documented, especially for natural resource stocks in less developed economies (Bailey, 1992; McKay and Acheson, 1987; Ostrom, 1990), and their details have been studied in many settings (e.g., Acheson, 1988; Dahlman, 1980, Eggertsson, 1992; Stevenson, 1991). Many writers on common property have noted the gains from group enforcement of rights to the resource (Ellickson, 1993; McKay and Acheson, 1987; Ostrom, 1990, and Stevenson, 1991), and we examine common ownership to take this empirical feature into account.[32]

Common property is best viewed as an intermediate case between open access and private ownership. Common property may arise out of explicit private contracting (e.g., unitized oil reservoirs, groundwater districts) or out of custom (e.g., common pastures and forests); it may have legal (e.g., riparian water rights) or regulatory (e.g., hunting and fishing regulations) bases that have implicit contractual origins. Contracting to form common property effectively creates a group that has exclusive rights to the resource (Eggertsson, 1992; Lueck, 1994). Acting together individuals can realize economies of enforcing exclusive rights to the asset. Equation (2.2) implies that waste can be reduced simply by restricting access to the asset.

A contracting model can illustrate how common property can limit waste from the rule of capture.[33] Contracting to form common property effectively creates a group that has exclusive rights to the resource. We assume that (contractual) agreement among group members pertains only to the group's size and the joint effort to exclude outsiders. In this setting, individuals acting together can realize economies of enforcing exclusive rights to the asset, so we also assume the costs of excluding (or policing) non-members can be represented as $p(n)$, where $p'(n) < 0$ and $p''(n) > 0$.

[31] Indeed Hardin's (1968) famous paper incorrectly characterizes the common pastures of English villages as open access resources when the historical record shows clearly that they were common property (e.g., Dahlman, 1980; Smith, 2000).

[32] Further evidence that common property regimes are productive is seen from the disasters that have occurred when they have been dismantled by the state (effectively creating open access) as in the forests of Nepal and Thailand (Ostrom, 1990).

[33] The model is based on Lueck (1994). Also see Caputo and Lueck (1994) and Wagner (1995). Others use evolutionary game theory models (e.g. Sethi and Somanathan, 1996).

A simple and customary method of allocating use of common property is a rule that grants equal access to all members of the group (Ostrom, 1990). Equal sharing of the asset avoids the explicit costs of measuring and enforcing individual effort (or use) but still creates an incentive for over use.[34] Effort is not explicitly part of the common property 'contract' so each member chooses his own effort (x_i) as he captures his share of the asset's output ($Y = f(\sum x_i)$ again) in competition with other group members. The size of the group is chosen to maximize the wealth of the group subject to the constraint of aggregate effort (X^c) by members operating in a common property regime, and in recognition of the costs of excluding outsiders. Optimal group size is a tradeoff between increased resource use with a larger group and increased enforcement costs associated with a smaller group. Formally the problem is

$$\max_n R = f\left(\sum_{i=1}^{n} x_i^c(n)\right) - \sum_{i=1}^{n}(w_i x_i^c(n)) - p(n), \tag{2.4}$$

where x_i^c is the individual's solution to the problem in equation (2.1). The optimal group size, n^c determines total effort[35] and must satisfy the first-order necessary condition

$$\frac{\partial R}{\partial n} = \sum_{i=1}^{n}\left[\left(f'\left(\sum_{i=1}^{n} x_i\right) - w_i\right)\right]\frac{dx_i^c}{dn} - p'(n) = 0. \tag{2.5}$$

Equation (2.5) states that the gain from an additional member in terms of a marginal reduction in policing costs must equal the marginal reduction in aggregate rent from overuse of the resource. The net present value of the common property resource is thus $V^c = \int_0^\infty R(x^c, t)e^{-rt}dt > 0$, where $V^* > V^c > V^{oa} = 0$. While the value of an asset governed by common property is less than its first-best value, it could have greater value than private property depending on the magnitude of the policing cost and overuse effects.

Dissipation from internal capture can be limited by maintaining a homogeneous membership. With equal sharing rules, a homogeneous membership maximizes the present value of a common property resource (Lueck, 1994, 1995b). Once a group chooses an equal sharing rule there is an incentive to maintain homogeneity. With heterogeneous members and equal shares, highly productive individuals will supply too little effort and the less productive will supply too much compared to a homogeneous equilibrium, so dissipation will increase. In effect, equal-sharing rules increase group wealth with homogeneity among group members. This provides an economic rationale for preserving homogeneity by screening potential members, by indoctrination, or by restricting the transfer of memberships.

[34] Common property might also be viewed as an output sharing contract with moral hazard. In this framework group members shirk as in a principal-agent model (see Lueck, 1994). Evidence of both types of common property—asset sharing (e.g., share access to a pasture) and output sharing (e.g. share the cheese produced from cattle on the common pasture)—are found in the empirical literature.

[35] Total effort is given by $X^c = \sum_{i=1}^{n^c} x_i^c$.

There are other potential limits on the capture behavior of individual common property owners that are not considered by the above model. For example, if group members expect to interact over long periods the incentive to overuse the resource may be limited by the desire to maintain a productive relationship. Accordingly, customary rules can evolve that restrict members, for instance, by limiting the size of private herds on a common pasture (Rose, 1986; Smith, 2000). For common resources that are attached to land such as oil, game, and water, ownership of the land can limit access to the resource. In effect, the group is the set of private landowners who have access to the common resource. In this case, private contracting to consolidate land holdings is a possible solution to the ownership problem for the attached resource (Libecap and Wiggins, 1984; Lueck, 1989).[36]

Another benefit of group ownership, besides internalizing externalities, is risk sharing.[37] Group ownership of land spreads the risk of uncertain events like crop failure, thereby providing a form of insurance (Ellickson, 1993). Group ownership also promotes egalitarianism, or equal sharing of output, which historically has been the motivation for various communal societies (Cosgel, Miceli, and Murray, 1997). But, as discussed above, these benefits must be weighed against the cost of group ownership in the form of diluted incentives for effort.

It is difficult to know how important common property regimes are in modern economies. Certainly families and other ‘close knit’ groups routinely use common property rights to govern resources. The ‘lobster gangs’ of Maine are perhaps the most famous case. A growing sector of the housing market is comprised of so-called ‘common interest communities,’ like condominiums, cooperatives, and homeowner’s associations, in which residents use a quasi-governmental structure for maintaining common areas (e.g., fitness rooms, pools, trails, open space) and providing certain local public goods (Dwyer and Menell, 1998: 807–887). (See the further discussion in section 7.5.) Common property seems to be less typical in business, perhaps because group ownership leads to costly transfers of rights that must ultimately be governed by political decision making. It may also be true that large-scale enforcement by the state (i.e., courts, police) has usurped the major advantage of common property. In water law, however, riparian water rights and the public trust doctrine (as we show in section 8.2.1 below) still contain important elements of a common property regime.

2.4. *State property rights*

A third, and increasingly important, category of property rights are those held by the state. Vast amounts of land, buildings, and capital equipment are owned by govern-

[36] On the other hand, there are problems when resource rights are tied to land ownership. For example, further parcelization of land can exacerbate the rule of capture as it has done with oil discovered in urban areas. In addition, linking rights can create incentives for further parcelization. For instance, riparian doctrine linking water to land sometimes yields long, narrow “bowling alley” parcels designed to extend water rights to many users (Dukeminier and Krier, 2002).

[37] Smith (2000) finds little support for the risk sharing thesis in the context of the medieval open fields system.

ments (local, state and federal). Local governments own schools, road ways, and fleets of police cars. States own universities, administrative buildings, and vast tracts of land, especially in the West, where statehood grants established state trust lands to be managed to finance schools. The federal government owns about one-third of the total land area in the United States, again with a much larger presence in the western states.[38] It owns the Outer Continental Shelf from the shore to the 200-mile international border and thus owns billions of dollars worth of oil–gas and other resources.[39] The federal government also has vast holdings of urban real estate (e.g., The White House, federal buildings throughout the country) and billions of dollars of capital equipment ranging from fighter jets and aircraft carriers to personal computers and desks.

The specific set of property rights that govern these state assets varies widely and has not been systematically analyzed by economists.[40] All are under the control of some administrative agency, be it the US Army, the state highway department, or the Bureau of Land Management. The statutes and regulations and political forces that govern these agencies vary widely and thus lead to a range of outcomes. Many federal lands are managed passively and are thus open access for many uses, especially for outdoor recreation such as cycling, fishing, hiking, hunting, and rafting. This is true for the bulk of land administered by the Bureau of Land Management, the Forest Service, and the Fish and Wildlife Service.[41] Other lands and uses are governed by a combination of price and non-price (lottery, waiting lists) mechanisms, but open to virtually all citizens in principle. Commercially valuable natural resources, such as coal, oil-gas, and timber, are routinely leased to private firms, who essentially have private rights over certain attributes of the land (Nelson, 1995). For example, ski resorts have long term leases to operate on federal lands, and commercial businesses such as hotels similarly tend to have long term agreements to operate in national parks. Moveable property like desks, planes and rifles are governed differently as well. In some cases state assets are assigned to individual users and thus become an almost exclusive usufruct right. It is well known that a soldier's rifle is 'his rifle' and no one else's. As a whole, the range of property rights regimes incorporate aspects of the three major types: private property, common property, and open access. Even so, what seems to be common to all of these regimes is a severe limit on transferability of rights, perhaps to limit the moral hazard incentives of agency bureaucrats.

Given the great variation in property rights, the analysis of state property not only requires a detailed knowledge of the asset and the relevant administrative agency but

[38] The total area of the 50 states is 2.27 billion acres and about 654 million acres (nearly 29%) is owned or administered by federal agencies. See Table 1-3, *Public Land Statistics 1999* (Bureau of Land Management, 1999).

[39] States tend to own the subsurface estate between the coast and the federal lands which begin from 3–5 miles out.

[40] The main exception to this is the large literature on marine fisheries where a myriad of administrative regimes define the rights to fish stocks (Munro and Scott, 1985).

[41] For these assets the typical rhetoric is that open access is good since 'they belong to all of us,' yet no one would make the same claim for an F-18 fighter jet.

also a workable theory of bureaucracy. The limited applicable literature is found in the analysis on natural resource agencies, especially those governing federal lands and marine fisheries. For instance, an early study by Stroup and Baden (1973) examines the behavior of the Forest Service and its management of national forests. They point out the different incentives faced by Forest Service managers compared to those of private forest owners and how interest groups influence agency behavior. Since that time there has developed a literature that has examined the economic efficiency of public land management. A common conclusion is that federal lands are not particularly well managed, and that these inefficiencies often are coupled with lower environmental quality (e.g., Hyde, 1981). More recently, Nelson (1995) notes the underlying and variable system of property rights in federal lands. Studies of the broadcast spectrum are also an example of economic analysis of property rights governed by administrative agencies (e.g., Hazlett, 1990, 1998; Hazlett and Munoz, 2004). Property rights to spectrum are, in fact, highly restrictive and generally non-transferable licenses to broadcast using certain technologies. These rights are similar to the rights to many public lands in their restrictions on use and transfer (e.g., federal grazing rights).

The relevant law for state property has origins in common law (e.g., mining on federal land in a first possession rule) but is primarily governed by statutes and regulations, all shaped by bureaucrats, interest groups, and politicians. These legal constraints shape the objective of agencies. For example, managers of state school trust lands in the West are typically mandated to maximize financial returns, and thus the land is managed intensively under a system of leases to private parties for uses ranging from farming to hunting to logging. National forests, however, are governed by federal 'multiple use' statutes which very often limit the ability of managers to generate revenue from forest use. These statutory constraints, in turn, shape the property rights that develop.

2.5. *Mixed property rights and complex assets*

Real property regimes are more complex than the open access, private property, common property, and state property discussions suggest. Real property rights regimes, in fact, are mixtures of these basic types. A rancher's land may seem to be private but this is only a partial description. The right to the grass for grazing is private but the streams running through the property may be open access for fishing or recreation; or the grass may be a lease from a federal agency with mineral rights held by yet another private party. The underlying oil reservoir may be governed by a unitization contract (subject to oversight by a state oil conservation agency) among many neighboring ranchers, essentially mimicking common property. Predator control for coyotes that roam across many ranches may likewise be governed by a common property regime. Similar scenarios are found in residential and commercial real estate, and Bailey (1992) found a mixture of ownership regimes among aboriginal peoples. The evidence, though far from systematic, suggests a mixture of rights. Because assets are a complex collection of valuable attributes, ownership is also a complex collection of rights (Barzel, 1982, 1997; Eggertsson, 1990; Ellickson, 1993; Kaplow and Shavell, 2002; Merrill and Smith, 2000; Rose, 1998; Stake, 1999), comprised of the four fundamental types.[42] One of the more distinctive complexities of rights—and one long recognized in law—is the distinction between use rights and transfer rights. The standard fee simple bundle included both the right to possess (use) and the right to transfer but many rights are only for exclusive use (e.g., riparian water rights, many servitudes).

Little work has been done to understand the forces that determine the optimal complexity of property rights.[43] This thus remains an important area for future work. Smith's (2000) study of the common field system of medieval Europe is one of the few to examine the economic logic of a mixed property regime. Smith notes that for crops the land in the typical village was private, but that for grazing the land was common property.[44] He notes how private property for crops provides incentives for investment and husbandry and how a larger scale of land ownership is optimal for grazing (of private herds). Lueck's (1989) study of wildlife law recognizes that wildlife is but one of many valuable attributes of the land for which the dominant property regime is most often governed by agricultural use. Ellickson (1993) similarly notes a wide range of mixed regimes, including legal and customary rights. These studies are important in furthering our understanding of the complexity of rights but are lacking a cohesive (and ultimately formalized) framework. The modern principal-agent literature on contracts, especially that on moral hazard, may be a starting point as our discussion of land leases in section 5 suggests. The major question is to what extent each individual attribute of an asset can be treated as an independent asset whose ownership is independently determined.

The common law of property can be said to begin with the *ad coelum* doctrine's mandate that ownership of land includes all attributes in an infinite projection above and below the earth's surface.[45] In this system the only ownership question is the size of the surface boundaries. The *ad coelum* framework ultimately breaks down because various attributes (as the rancher example shows) have different surface projections. Thus the optimal tract size of land for a home may be one acre, but for an oil reservoir it may be ten thousand acres and for an airshed it may be much larger still. The law has long recognized the limits of the *ad coelum* doctrine and has developed to accommodate the demand of different attributes of land. The law of servitudes and the law of separate estates in water and minerals are clear examples. Modern public administration of environmental resources is a recent application (e.g. Rose, 1998). The law of nuisance and trespass, the focus of Coase's analysis, has to do with conflicts that ultimately arise between the owners of adjacent parcels, which derive from complex assets with various dimensions of use. The doctrine of private necessity, for example, is an exception to the

[42] Rose (1998, p. 96) calls un-enforced attributes of land by the term 'unpropertized common resources'.

[43] Karpoff's (1987) study of fisheries regulations, which we note later, implicitly recognizes complex assets.

[44] Smith labels this regime a 'semi-commons,' though like the term anti-commons it is not clear that this identifies a distinct and fundamental economic regime.

[45] The complete Latin phrase is *cujus est solum ejus est usque ad coelum*, which translates 'to whomsoever the soil belongs, he owns also to the sky and to the depths' (Dukeminier and Krier, 2002, p. 141).

law of trespass, which actually allows one to use another's property in an emergency.[46] Thus emerges the traditional legal concept of property rights as a 'bundle of sticks;' an idea that accurately meshes with the complexity and mixed ownership of real assets. As Ellickson (1993) notes, the common law allows a wide variety of subdivision of rights in time, use, and space.[47]

3. The origin of property rights

This section examines the origin of property rights. In both custom and law first possession has been the dominant method of establishing rights, and the rationale for and the effects of this mechanism will be examined closely. It will be clear that the manner by which possession is defined and enforced will be crucial in the type of rights that are created. Alternatives to first possession are also examined including auctions, lotteries, and administrative assignments.

3.1. First possession

First possession rules can operate on different margins. For instance, the rule can grant ownership of a single bison to the first person that kills it under the so-called rule of capture, or it can grant ownership of the entire herd to the first person that claims ownership of the entire living herd. The behavior of the possessor and the use of the bison resource will obviously differ in the two cases. In the initial case, first possession applies to the *flow* of output from the *stock* of living bison, while in the second case the rule applies to the stock itself. In the bison example, the rule of capture is expected to emerge—and in fact did—because the cost of enforcing possession to the live herd is prohibitive (Lueck, 2002).

Figure 3.1 illustrates the effects of a first possession rule, beginning with an unowned asset. As the left branch of the figure shows, if applied to a stock, private property rights are established directly through possession. On the right branch, if only a flow (or a portion of the stock) can be possessed, the rule of capture ensues. Thus both paths have the potential for dissipation, either from a race to claim the stock or from open access exploitation. In a race, dissipation takes the form of excessive investment prior to ownership, but the resource is unaltered. In contrast, under the rule of capture, dissipation manifests as overuse of the resource (and possibly damage).

[46] *Ploof v. Putnam* 81 Vt. 471, 71 A 188 (Supreme Court of Vermont, 1908). Here a person was allowed to tie-down a boat at a private dock during a severe storm without permission. The court ordered the user to pay for the use of the asset.

[47] Some recent studies have noted that property law tends to allow only a certain number of standard bundles of rights (Ellickson, 1993; Merrill and Smith, 2000; Hansmann and Kraakman, 2002). The explanations for this *numerus clauues* doctrine focus on measurement and information costs.

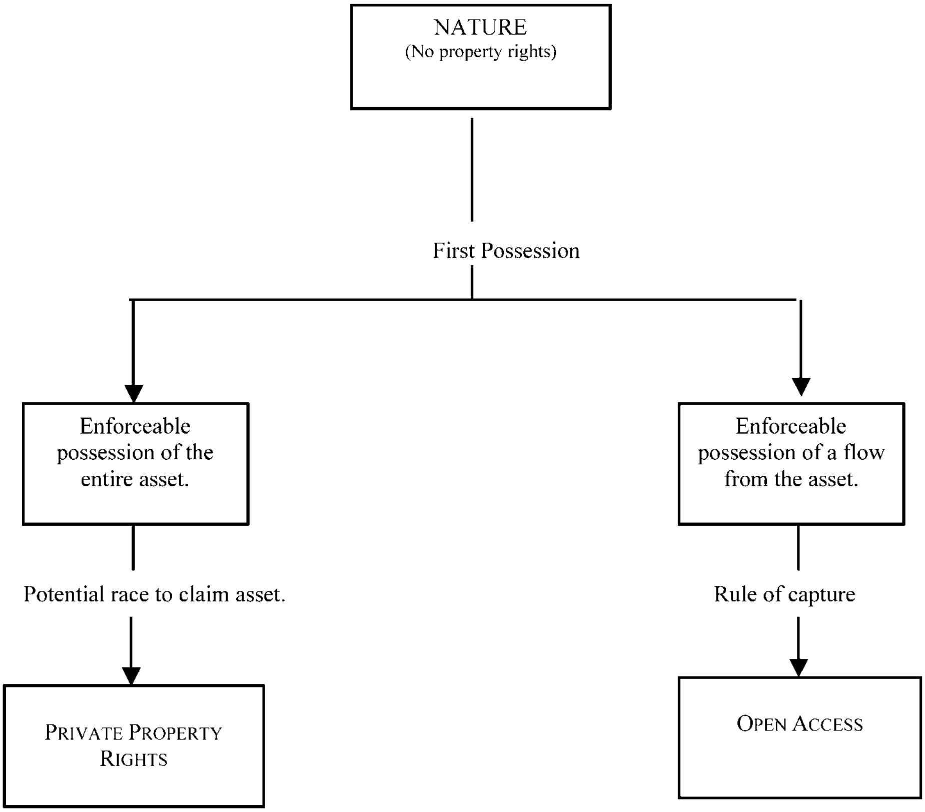
NATURE
(No property rights)
First Possession
Enforceable possession of the entire asset.
Enforceable possession of a flow from the asset.
Potential race to claim asset.
Rule of capture
PRIVATE PROPERTY RIGHTS
OPEN ACCESS

Figure 3.1. Property rights under the rule of first possession.

The stock-flow distinction also illuminates the temporal dimensions of ownership. For example, possession could grant ownership of a pasture in perpetuity or it could simply grant ownership of the grass currently being grazed by one's livestock. Perpetual ownership means ownership of the stock, while a shorter term of ownership means ownership of some flows. Granting rights to stocks also confers ownership to the future stream of flows, so the formal economic model is inter-temporal. Granting rights to flows, on the other hand, means ownership is a one-time event, so the formal economic model examines just one period. Of course, there can be ownership rights of intermediate term (e.g., patents, copyrights) but this simple dichotomy covers most of the important cases and serves to clarify the model.

Consider an asset (e.g., a plot of land) that yields an instantaneous (net) flow of benefits $R(x(t))$, where $x(t)$ is the amount of a variable input supplied by private owners at time t.[48] Let r be the interest rate, and assume the flow value, $R(t)$, grows over time at the continuous rate $g < r$, so that the value of the asset grows over time. Also assume that each period's return is independent of past returns.[49] The term g can be thought to measure increases in the demand for the asset, perhaps because of population growth. This formulation also recognizes the usual case that during early periods assets are not sufficiently valuable to cover the costs of establishing ownership. The first-best, full-information outcome is

$$V^{FB} = \int_{t=0}^{\infty} R(x^*(t))e^{-(r-g)t}dt, \tag{3.1}$$

where $x^*(t)$ is the optimal input level in period t and the first-best time to begin use of the asset is $t = 0$. In general, V^{FB} is not attainable because of the costs of both establishing and enforcing rights that efficiently allocate use of the resource.

3.2. *Claiming the asset*

The left-hand side of Figure 3.1 shows the case when ownership of the asset is granted to the first person to obtain possession of the entire stock. To simplify, we assume that the method of possession does not damage other resources; this is equivalent to assuming that the asset is a simple, single attribute good. The first claimant thus obtains exclusive rights, into the indefinite future, to the flow of rents, $\int_0^\infty R^*(t)dt$, generated by the asset. Since establishing a bona fide claim will be costly and because $g < r$, rights may not be worth enforcing. Under these conditions, property rights to the asset will emerge, after an initial period without ownership, as the value of the asset increases (Demsetz, 1967). Maximizing resource value is, in effect, a problem of optimally timing the establishment of rights under first possession.

Now assume there are one-time costs, C, of establishing enforceable rights, or demonstrating possession which grant the owner the exclusive right to the stream of production for all time. If there is a single potential claimant, the flow from the asset (and the rents) is available after rights to the stock are established. The decision to claim the stock is the result of private maximization which, in this case, means the net present value of the asset is

$$V^S = \int_{t^s}^{\infty} [R(x^*(t))e^{-(r-g)t}dt] - Ce^{-rt^S}, \tag{3.2}$$

where t^S is the time at which ownership of the stock (and the flow of output) is established under first possession. The optimal time to establish ownership is when the marginal return from waiting, given by the present value of the asset's flow at t^S, equals

[48] The model here is derived from Lueck (1995b, 1998).

[49] One can think of $R(t)$ as the steady-state flow of benefits. Note that if $g > r$, the present value of the asset would be infinite.

the marginal cost of waiting, given by the present value of the opportunity cost of establishing rights at t^S, or $R^* e^{-(r-g)t^S} = rCe^{-rt^S}$. Inspection of (3.1) and (3.2) shows that the value of the asset clearly falls short of first-best, or $V^* < V^{FB}$. This is because the net value of the asset must now account for the costs of establishing ownership, and the fact that these costs delay ownership and production to t^S from $t = 0$.

A first possession rule can dissipate value when there is unconstrained competition among potential claimants.[50] In the simplest case with homogeneous competitors, potential claimants gain ownership by establishing possession just before their competitors. A claim is worth staking as long as the net value of the asset is positive, so a competitive rush to claim rights causes ownership to be established at exactly the time, t^R, when the present value of the rental flow at t^R equals the present value of the entire costs of establishing ownership at t^R, or when $R^* e^{-(r-g)t^R}/(r-g) = Ce^{-rt^R}$. In such a race, rights are established prematurely at t^R, where $t^R < t^S$.[51] More important, the race equilibrium implies that the rental stream is fully dissipated; that is,

$$V^R = \int_{t^R}^{\infty} [R(x^*(t))e^{-(r-g)t} dt] - Ce^{-rt^R} = 0. \tag{3.3}$$

Heterogeneity among potential claimants can reduce, or even eliminate, the dissipation of wealth (Barzel, 1994; Lueck, 1995b).[52] Assume there are just two competitors (i and j) for ownership of the asset with possession costs $C_i < C_j$. Also assume that neither party knows each other's costs. In a race, person i gains ownership just before the closest competitor makes a claim, at time, $t^i = t^R - \varepsilon$, and earns rent equal to the present discounted value of his cost advantage, V^{R_i}. The key implication is that as the heterogeneity of claimants $(C_j - C_i)$ increases, the level of dissipation will decrease. The analysis remains the same with rental value differentials such as $R_i \neq R_j$ or different expectations about the rate of growth of the flow value, $g_i \neq g_j$. In the extreme case, where just one person has costs less than the net present value of the asset's flows, the first-best outcome is achieved. Since only one person enters the race, there is no dissipation.

Altering the assumption about information can alter the racing equilibrium. Fuedenberg et al. (1983) and Harris and Vickers (1985) show that if competitors have complete information about each other's talents a race will not ensue because only the low-cost individual will have a positive expected payoff of entering the race; that is, V^S is achieved if $C_i < C_j, i \neq j = 1, \ldots n$.

Even though claimant heterogeneity can limit or eliminate racing dissipation, there arises the possibility that a claimant can gain a cost advantage by expending resources,

[50] This phenomenon was first studied by Barzel (1968) in the context of research and technological development. Also see Mortensen (1982).

[51] The single claimant solution yields $t^S = (\ln r + \ln C - \ln R)/g$ while the race model gives $t^R = (\ln(r - g) + \ln C - \ln R)/g$. Inspection reveals $t^R < t^S$.

[52] Suen (1989) also notes the importance of heterogeneity in reducing dissipation in the context of rationing by waiting.

thereby altering the margins of dissipation (McFetridge and Smith, 1980). For example, if competing claimants can acquire the technology to achieve the minimum costs (C_i), then homogeneity and the full dissipation equilibrium is re-established. This extreme result, however, relies on the assumption that homogeneity can be attained easily by investing in the low cost claimant's technology. The more likely reality is that claiming costs depend not only on endogenous investment decisions but also on exogenous forces that generate and preserve heterogeneity. Consider two possibilities. First, if the distribution of talent across individuals is not equal, some people will have innate advantages that will be difficult or impossible to overcome with investment. Second, if there is random variability in opportunities, then some individuals will be in the position of being the low cost claimant; again, investment is unlikely to destroy the random advantage.

Because first possession is a rule that restricts competition to a time dimension, there is another reason why investment cannot routinely eliminate heterogeneity. Cost advantages, no matter how they were gained initially, are expected to diminish over time because potential investors ultimately will gain information that allows them to mimic the behavior of the low cost person (Kitch, 1977; Suen, 1989). As long as costs depend on exogenous factors, dissipation will be incomplete. In the worst-case race equilibrium, the first claimant will own just the value of his exogenous advantage; in the best-case, extreme heterogeneity or the full information game theory equilibrium, the first claimant will own the full potential value, V^S, of the asset.

3.3. The rule of capture for asset flows

When the costs of enforcing a claim to the asset are prohibitive, ownership can be established only by capturing or 'reducing to possession' a flow from the asset. (See the right side of Fig. 3.1.) The rule of capture—simply a derivation of the rule of first possession—will occur when enforcing possession of the flow is cheaper than enforcing possession of the stock. Wildlife and crude oil are the classic examples: ownership is established only when a hunter bags a pheasant or when a barrel of oil is brought to the surface. The stock itself, be it the pheasant population or the entire underground reservoir of oil, remains unowned. As a result, the new 'race' is to claim the present flow $R(t)$ by capturing the product (e.g., the dead pheasant) first.

As a rule of capture, first possession can lead to classic open access dissipation (Epstein, 1986; Lueck, 1995a, 1995b). Under the rule of capture no one owns the asset's entire stream of flows, $\int_0^\infty R(t)dt$. Now the formal economic analysis of dissipation is just one-period, rather than inter-temporal as in the race, and in fact is identical to the open access model developed in section 2.1, with an equilibrium level of effort determined by equation (2.2).

3.4. First possession in law

The law of first possession seems to be consistent with the model that includes two potential paths of dissipation (racing and over-exploitation). When first possession has

Table 3.1
First possession rules

Asset	Possession rule	Stock-flow & duration of rights
Chattels (abandoned, lost, unclaimed)	recover or show intent to recover	stock—permanent
Intellectual property	invent, write	stock—varies (17–100 years)
Land	occupation & cultivation of land	stock—permanent
Minerals (hard rock)	locate mineral deposit	stock—permanent
Ocean fisheries	land fish	flow—current catch
Petroleum	bring oil to surface	flow—current production
Water—appropriation doctrine	develop a diversion plan	stock—permanent
Water—riparian doctrine	pump or divert water	flow—current use
Wild game	kill or capture animal	flow—current kill

the potential for a race, the law tends to mitigate dissipation by assigning possession when claimant heterogeneity is greatest. On the other hand, when first possession breeds a rule of capture, the law tends to limit access and restrict the transfer of access rights to limit open access exploitation. It should be noted that judicial opinions and statutes may use such terms as 'first in time, first in right,' 'priority in time,' or the 'rule of capture.' Regardless of the precise legal terminology, all of the subjects examined below are governed by rules in which legitimate ownership is created by establishing possession before anyone else. Table 3.1 summarizes some important first possession rules.[53]

In those cases where first possession rules establish ownership in a resource stock, a number of common principles are evident. First, possession tends to be defined so that valid claims are made at low cost and before dissipating races begin, thus exploiting claimant heterogeneity.[54] Second, once rights are established, the transfer of rights to the resource is allowed routinely. Third, the use of auctions or other administrative allocation mechanisms are high cost alternatives.

In certain cases, establishing possession of an entire stock is especially costly and leads to the rule of capture, as in the case of so-called 'fugitive' resources (Rose, 1986) such as oil and wildlife. In these cases a number of common principles can be found. First, the rule of capture may not produce severe dissipation when there are but a few users or when there are what Rose (1986) calls 'plenteous' goods. Thus open access may persist optimally as in the case of nineteenth-century whaling. Second, when

[53] The analysis here suggests broad confirmation of the economics models, but the literature shows considerable disagreement among law and economics scholars on the merits of first possession rules (Merrill, 1986). For instance, in studies of homesteading (Anderson and Hill, 1990) and water (Williams, 1983) first possession has been criticized as causing wasteful races. In contrast, studies of the broadcast spectrum (Hazlett, 1990), homesteading (Allen, 1991), and patents and mining (Kitch, 1977) argue that racing dissipation is minimal.

[54] In a rare empirical study on this issue, Lerner (1997) finds evidence consistent with racing in the U.S. disk drive industry.

dissipation becomes severe, access to the resource tends to be limited through legal, contractual, or regulatory methods. Third, transfer of rights to capturable flows tends to be restricted. Contemporary fishing regulations are perhaps the best example of such regimes, though the complexity of assets makes it difficult to eliminate the margins for dissipation (Karpoff, 1987).

Even possession under the rule of capture can vary, as illustrated in the famous case of *Pierson v. Post* where the court was divided over whether possession of a wild fox was determined by "hot pursuit" or physical capture.[55] A similar distinction was present in nineteenth century Atlantic whaling (Holmes, 1881; Ellickson, 1989). Here, the rule of capture typically required that a harpoon be fixed to the mammal before a legitimate ownership interest was established, the 'fast-fish, loose-fish' rule. In the case of the aggressive sperm whale, however, the 'iron holds the whale' rule granted ownership to a whaler whose harpoon first was affixed to the whale so long as the whaler remained in fresh pursuit. The law seems to recognize how the precise way in which possession is defined will influence the outcome and tends to define possession so that waste (e.g., fruitless or dangerous whaling effort) will be minimized.

What must be done to maintain a legitimate claim? Ownership, says Blackstone, remains with the original taker, 'till such time as he does some other act which shows an intention to abandon it.'[56] In general the law tends not to require a claimant to continually exert the effort required for an initial claim, but he cannot remain an owner without incurring some continued possession costs (Holmes, 1881).[57] An owner must actively and continuously enforce his ownership claim, regardless of whether he obtained ownership by first possession or by subsequent method such as purchase, inheritance, or bankruptcy. The law has two responses to a party lax in exerting effort at continued possession. If an owner intentionally ignores the property it can become abandoned and subject to being reclaimed under first possession. In certain cases, (e.g., minerals, trademarks, water) specific rules, often lumped together as 'use-it-or-lose-it,' have developed to determine precisely when the right has been abandoned. If an owner is simply inattentive enough to allow another party to establish continued use of the property, then adverse users can ultimately gain ownership under the doctrine of adverse possession

[55] *Pierson v. Post*, 1805, 3 Cal. R. 175, 2 Am. Dec. 264 (Sup. Ct. of N.Y. 1805). Dharmapala and Pitchford (2002) develop a formal model of the differential incentives under the two rules.

[56] Book II, Chapter 1. Of course, property rights can also be relinquished by gift or sale to another.

[57] This principle is clearly articulated in the famous 'dung case,' *Haslem v. Lockwood*, 37 Conn. 500 (1871). In *Haslem* the plaintiff was a farmer who gathered manure from the ditch along a public highway into 'heaps,' leaving them overnight while he returned to his farm to get a cart for transport of the heaps. Before he returned the defendant had begun to load the heaps and take them away. The court, in deciding for the plaintiff, ruled that the manure was abandoned property in the public ditch, that the plaintiff established ownership via first possession by piling the dung into heaps, and finally, that the plaintiff having established ownership did not have to exert the same effort to maintain possession and was therefore justified in returning home to fetch his cart. Implicit in this case and elsewhere is the fact that collective institutions (e.g., courts, custom, police) actively enforce property rights once they are established, thus minimizing the resources devoted to continued possession.

(see section 6.1). Thus the law requires that an owner continue to exert effort to maintain possession but certainly not to the degree initially required to establish possession. In Holmes's words (1881, p. 236): 'Everyone agrees that it is not necessary to have always present power over the thing, otherwise one could only possess what was under his hand.' The general rule of not requiring the same effort for continuing possession as for establishing possession recognizes economies of enforcement by collective institutions and a protection of specific investments by the original claimant.[58]

A first possession rule that leads to an optimal system of ownership for one attribute can leave rights unspecified to another attribute.[59] Establishing rights to land for farming, for instance, might create a system of rights inconsistent with the optimal use of wildlife or groundwater. The process of establishing possession might cause damage to adjacent environmental assets, as when the diversion of water under prior appropriation damages in-stream resources (Leshy, 1987; Sprankling, 1996). Indeed, the application of first possession to environmental goods (e.g., scenic view) is not well developed in the law. Private contracting to consolidate land holdings is a possible solution to the ownership problem for the attached resource, but this is an imperfect solution when contracting costs are positive (Libecap and Wiggins, 1984; Libecap, 1989). For example, detailed property rights to small, urban parcels of land can lead to severe open access dissipation for subsurface oil and gas production. Recent work by Hansen and Libecap (2004) shows that soil erosion during the 'Dust Bowl' can be similarly viewed as a failure of private contracting among many small farmers.

Possession rules can also swing dramatically from a rule of capture to a perpetual right to a stock. Water law illustrates the issue clearly. Under absolute ownership a landowner can claim groundwater under the rule of capture by pumping water to the surface; under prior appropriation, however, a successful first claimant earns a permanent withdrawal right to a measured quantity extracted each year. Indeed, such a switch in regimes begs the question of what is the actual stock that is valuable to potential users. Is the bison herd the valuable stock, or is a single bison (which can yield meat and hides) a valuable stock in its own right? Ultimately the answer depends on the uses of the resource as well as on the relative costs (e.g., claiming possession, enforcing common property).

First possession rules are still relevant and likely to be important in the future. Berger (1985) notes many cases not examined here where first possession is the primary rule. For example, while the common law has tended to move away from the 'coming to the nuisance' doctrine (Wittman, 1980; Pitchford and Snyder, 2003), nearly all states have enacted 'right to farm' statutes which effectively codify this first possession principle; namely, that no one can make a legitimate nuisance claim for activities in place prior to

[58] Continued possession or maintenance costs can be added to the first possession model, noting that net rents are $R(t) - c(t)$ where $c(t)$ is the current cost of maintaining possession. This addition will increase t^* in the claim model.

[59] Nuisance law, as discussed in section 7.4, addresses these problems.

a location decision by the affected party (Berger, 1985). The recent environmental policies that use transferable use or access permits require an initial allocation of property rights. For both fisheries regulations that use individual transferable quotas (ITQs) and pollution emission systems with transferable permits, rights tend to be established by being grandfathered in to the permit system. For fisheries, allocations have been based on historical catch; for pollution, allocation has been based on historical emissions (e.g., the sulfur dioxide trading program under the Clean Air Act amendments of 1990). Some economists have considered this a 'free distribution' (e.g., Stavins, 1995) or 'give away,' but it is alternatively viewed as an allocation based on first possession. In these cases, first possession may protect the specific investments made by the original users of the assets and avoid the administrative and rent-seeking costs of auctions. Though it might seem reasonable to think that the era of discovering new resources has long passed, space and the deep sea may have surprises to offer. In space, geosynchronous satellite orbits have been claimed by first possessors, but the deep sea has been treated differently. For example, Epstein (1979) noted that the Law of the Sea conference rejected first possession rules for allocating claims to deep sea minerals, while recent legislation awards ownership of abandoned shipwrecks found in U.S. waters to the federal government rather than to the finder (Hallwood and Miceli, 2006).

3.5. Alternatives: auctions, bureaucracy, politics, and violence

Law and economics scholars studying first possession have often recommended auctions as the efficient method of establishing rights without closely examining the costs of auctions (e.g., Barzel, 1968; Coase, 1959; Haddock, 1986; Posner, 2003; Williams, 1983). Assuming the same costs of establishing the rights (C), the winner of the 'ideal' auction pays V^S and begins production at t^S, thus maximizing the value of the asset. Yet, in practice, auctions will entail real and often large costs (Epstein, 1979; McMillan, 1994). Under first possession, private claimants must bear the cost, Ce^{-rt}, of enforcing a claim to the resource. Similarly, before the auction can take place, the state must establish rights to the asset at a cost, C^Se^{-rt}, and also incur costs, C^Ae^{-rt}, of administering the auction. In addition, the state must survey and police the resource, determine what size parcels of the asset to sell, the method of auction to use, and so on (McMillan, 1994). If the state cannot protect property rights adequately after the auction, potential buyers will bid less than V^*.[60]

Epstein (1979) also notes that interest groups will attempt to alter the auction rules to suit their own advantage, leading to further dissipation of rent. Indeed, he notes that administrative alternatives simply were not available (*i.e.*, too costly) during much of the development of the common law. As a result, only if the state's costs $(C^S + C^A)e^{-rt}$

[60] Allen (1991) argues that this is the reason homesteading was often chosen over auctions for assigning rights to frontier land in many countries.

are less than Ce^{-rt} will V^* result from an auction.[61] The choice between auctions (or other administrative policies) and first possession is ultimately a trade-off between costly auctions and potential dissipation from races. In some cases—future patentable innovations, sunken treasure, and the unused electromagnetic spectrum—the resource cannot be auctioned simply because it has yet to be identified. First possession rules, as Epstein (1979) notes, generate incentives for private parties to discover new resources and new dimensions of known resources.

4. The evolution of property rights

To this point several different property rights regimes have been studied in isolation, and the establishment of rights has been considered under the rule of first possession. This section examines the determinants of changes in property rights and how these changes take place. Though changes or differences in property rights can be examined with cross section or time series data, the earliest studies focused on temporal changes, and thus the term "evolution of property rights" has come to define the literature.[62]

4.1. The Demsetz thesis

The evolution of property rights is one of the oldest topics in the economics of property rights, beginning with Demsetz's (1967) pioneering paper.[63] Demsetz argues that property rights emerge to internalize the externalities present in open access.[64] Further, in what has become the classic argument on the topic, he posits that an increase in the value of an asset will increase the gains from ownership and thus lead to the creation of property rights. In support of this thesis, Demsetz recounts the anthropological evidence of alterations in property rights among the Montagne Indians of Quebec during the 18th century. Prior to the emergence of the beaver trade with Europeans, rights to beaver could be characterized as open access. However, once the trade increased their value, property rights to beaver populations emerged and were held by family units. The story of the emergence of rights to beaver among the Montagne has become the most famous story in the economics of property rights.[65]

[61] Hazlett and Munoz (2004) argue that the time at which a resource is used can be substantially delayed under an auction system (the choice of t in the model) because of lobbying during the design period. They further estimate that these delay costs were around $9 billion in the case of recent spectrum auctions in the United Kingdom.

[62] Anderson and Hill (1975) appear to be the originators of this phrase. Recently the Journal of Legal Studies (Vol. 31, No. 2 (part 2) (June 2002)) published a special issue titled 'The Evolution of Property Rights.'

[63] Although Demzetz's paper might seem to fit in the 'origins' topic in the previous section, it is more appropriately viewed as a study of the choice among various property regimes, so we place it here.

[64] Demsetz actually uses the term 'common property' to describe what is now called open access.

[65] But see McManus's (1975) account of property rights to beaver which indicates a more complex system of rights than the all-or-none choice Demsetz describes.

4.2. Empirical studies

The Demsetz thesis was not again explored until Anderson and Hill's (1975) study of the emergence of property rights to rangeland, livestock, and water in the American West. Anderson and Hill argue that the history of the West is largely consistent with Demsetz's thesis; as the frontier was settled assets became more valuable, and property rights emerged out of what we would now call open access. In a remarkably convincing historical analysis they show how the range was privatized after the introduction of barbed wire dramatically reduced the cost of enforcing rights to grasslands. This history shows how, holding resource values constant, changes in property rights enforcement costs can have dramatic affects on the choice of property rights regimes.[66]

Umbeck (1977) and Libecap (1978) study the establishment of rights to gold and silver fields in California and Nevada, respectively, and find a history that again corresponds to Demsetz's beaver. In fact, the California gold rush is an even better application than the beaver because the discovery of gold signified a sharp increase in the asset's value and the property system that developed was much more detailed than that developed by the Montagne. Furthermore, for the gold case, there was no preexisting society as in the Montagne case; open access truly was the prior regime. In his sweeping study of economic history, North (1981) suggests that the general rise of agricultural societies, with private property rights in land, is consistent with this view of emerging rights. Indeed, one might argue that the settlement of North America is broadly consistent as well.[67] Over time rights to land, water, minerals, and even air in recent times, have been established as asset values have increased.

Econometric evidence confronting the Demsetz thesis has been scarce because of the severe data requirements. Such a test requires data on property rights regimes and the relevant economic parameters. Quantification of property regimes is particularly difficult and over a time series even harder. Libecap (1978), however, couples his historical account of changes in mining law with some econometric evidence, showing in a short time series that mining law became more precise as the value of mineral deposits increased. More recently Geddes and Lueck (2002) use panel data on state laws defining the rights of married women to hold property and contract, and find that states with a greater potential value of human capital (as approximated by levels of wealth, education, and the size of the market) tended to be the first states to expand rights for women. Geddes and Lueck's study is consistent with Demsetz and also with Schultz (1968) who noted that individual freedoms (or rights to one's own human capital) have tended to increase with increases in the value of human capital.

Despite numerous studies in support of the Demsetz thesis, there are many instances where property rights did not emerge even as asset values increased considerably. The

[66] What is not discussed in Anderson and Hill is the destruction of Native American property regimes as these asset values increased.

[67] Eggertsson (1990) summarizes this literature; see also the Journal of Legal Studies (2002).

case of oil and gas is perhaps the most dramatic, where rights to underground reservoirs remained subject to a rule of capture (see section 3.1) even as the value of these resources rose dramatically (Libecap, 1989; Libecap and Smith, 2002). Rights to the oil and gas stocks themselves took nearly a century to develop and never emerged in common law doctrine. In another example, property rights never emerged for the wild bison herds despite the rather dramatic increase in the market value of the bison with the advent of the bison hide market (Lueck, 2002). In fact, it is precisely during the period of the most intense market activity that the bison's demise was swiftest. Property rights to marine fisheries often have also remained open access for extended periods, despite significant increases in the asset's value.[68]

4.3. The theory of rights evolution and variation

Demsetz's original theory was informal and simple: increases in the net benefits of enforcing rights would increase the level of rights enforcement. His main theoretical contribution was simply and importantly to note that property rights themselves are economic goods amenable to the tools of economic theory and potentially subject to empirical analysis. While it seems trivial now, as do many breakthroughs, the insight has been critical to the economics of property rights and institutions. Yet, Demsetz did not develop a formal model, and his paper said little about the costs of a property regime, the mechanism of choosing rights, and the form of property rights.

Umbeck (1977) was the first to formalize Demsetz in his study of the California gold rush. Umbeck assumes that the net value of rights (V) was simply the benefits of rights (B) less the cost of enforcing rights (C). He assumes further that B was exactly the market value R of the asset (e.g., the price of a beaver pelt, or the rental value of a plot of land), and that enforcement costs positively depended on the asset value $C = c_0 + c(R)$ where $c'(R) > 0$. The first-best, or zero transaction cost value of the asset is $V^* = R$, but the second-best value of the asset is $V = R - C(R)$. Property rights emerge only when $R > C(R)$, so that for low asset values the asset remains unowned. This is exactly the Demsetz thesis.[69]

Implicit in the Demsetz model was the assumption that the there will exist an asset value for which $R = C$, so that there are values for which property rights will be enforced and values for which they will not. Umbeck, however, notes that this outcome depends on the structure of the enforcement cost function C.[70] It is possible that as asset values increase there may be an even greater incentive to steal the asset thus raising enforcement costs. Simply, if enforcement costs rise faster than asset values, then the

[68] This should not be taken as evidence against Demsetz's general thesis, which is simply that property rights respond to economic costs and benefits.

[69] The first possession model from section 3.1 also implies that rights will emerge over time as asset values increase, given some costs of claiming and a first possession rule.

[70] Field (1989) also notes that enforcement costs depend on asset values. His model focuses on the number of owners of a tract of land, or what he calls the 'optimal number of commons.'

implication is that no property rights will be established at all, regardless of how high the asset value becomes. This means that if $c'(R) > 1$,[71] then no rights will be established because $C(R) > R$ for all values of R. This means the only clear prediction from the model is that parametric decreases in enforcement costs will increase the probability that property rights will emerge.[72] Thus changes in asset values do not give unambiguous predictions. Allen (2002) notes the possibility that enforcement costs are increasing and convex in asset values (i.e., $c'(R) > 0$, $c''(R) > 0$). This extension implies that at lower asset values, an increase would lead to the establishment of property rights, but that at higher asset values, a further increase could actually lead to a reversal or abandonment of property rights.[73] A consideration of complex assets can also alter the model (Lueck, 2002). If, for instance, land is valuable for the production of both bison and wheat, then an increase in the value of bison might not lead to an increase in rights to bison if this increase is correlated with an increase in the value of wheat, which requires land ownership on a smaller scale than is optimal for bison.

The choice of property rights can be put in a framework in which the maximization problem is $\max\{V_1, \ldots, V_n\}$ where V_i is the net value of the asset generated under the i^{th} property rights regime (e.g., common property, state property, private property).[74] Each regime's value depends on market parameters and transaction cost parameters. With multiple viable choices an analytical solution may be not be available, but in many empirical settings the choices may be rather limited. In the case of just two alternative property regimes, comparative statics predictions can be generated from $\partial(V^i/V^j)/\partial\chi$, where V^i and V^j are value functions for property regimes i and j and χ is a parameter. If this derivative can be signed, then there is a prediction for the choice of property regime.

4.4. The mechanism of rights changes

The analysis of the mechanism by which property rights are established can be divided into several categories. First, there is what might be called an 'institutional invisible hand,' which can be attributed to Demsetz and even to Coase (1960). This is often linked to the common law (Posner, 2003). The evidence on whether the law has evolved in a manner consistent with efficient property rights is mixed.[75] The development of the prior appropriation water doctrine in the western states seems to be an affirmative case. In one of the defining cases, *Coffin et al., v. Left Hand Ditch Co.*,[76] a Colorado court

[71] Since $B = R$, $B' = 1$.

[72] The case of the barbed wire fence fits this prediction (Anderson and Hill, 1975).

[73] Such reversals have been noted by Anderson and Hill (1975) and Smith (2002).

[74] Under the conditions of the Coase Theorem, of course, each regime would generate identical asset use and value.

[75] There is limited common law doctrine that compels adjoining landowners to share costs of common assets (e.g., walls as in *Day v. Caton*, 119 Mass. 513 (1876) and groundwater drainage as in *Ulmer v. Farnsworth*, 15 A. 65 Maine (1888)). See Ellickson and Been (2000).

[76] 6 Colo. 443 (1882).

noted that the riparian system used in eastern states was not useful for western water, which needed to be diverted for use. Yet, in recent years courts in western states have been reluctant to allow water rights to be defined for increasingly valuable 'instream uses' such as those for recreation and wildlife.[77]

Second, game theory suggests that in the presence of repeated interaction, agents in open access can generate conventions or norms in which the parties agree to create a system of rights.[78] For instance, Sugden (1986) suggests an evolutionary explanation for such property conventions as 'first possession' and respect for property rights. Experimental work by Walker, Gardner and Ostrom (1990) shows how property regimes can emerge out of open access with repeated interaction, even with anonymous players.

A third line of analysis, closely related to the second, is contracting for rights (Libecap, 1989). That is, when the gains from another ownership regime exist, there is the potential for existing users or those who have access to form a deal to establish a new regime. Such an outcome is explicit in the formation of a unitization agreement to establish rights to an underground oil reservoir among parties who previously operated under a rule of capture. For such a contract to be an economic equilibrium there has to be rent from the new property regime, and each party to the contract must expect to increase their own rents. In the language of modern contract theory, a successful contract must satisfy the incentive compatibility and individual rationality constraints of all parties. Libecap (1989) finds that in many cases there is sufficient heterogeneity and information asymmetry among contracting parties that it is prohibitively costly to find a contract that meets the individual rationality constraints of all. Thus open access under a rule of capture can persist even when the potential rents are huge.

A fourth mechanism by which rights can be established is through politics and statutory rule-making.[79] Since rights are often initiated via political institutions, there must be rents for the political actors (e.g., politicians, interest groups, and bureaucrats) to implement the changes. Thus, there is yet another set of incentive compatibility and individual rationality constraints to add to the purely private contracting model. Rose (1998) recognizes these forces and, without developing a model, suggests that the modern evolution of rights to environmental goods has been more or less consistent with the Demsetz thesis. Riker and Sened (1991) explicitly show how transferable rights to airport landing slots—established in the 1980s—did not emerge until the Secretary of Transportation and the Office of Management and Budget signed on. More generally, Levmore (2002) examines some interest group explanations for property rights changes and formation.

[77] Only recent statutory changes have allowed the definition of these rights, suggesting that there is a tradeoff between common law and statutory rule-making. Similar outcomes can be found in the law of oil and gas, wildlife, and groundwater.

[78] Greif (2006) makes a similar argument for other self-enforcing institutions (or social rules generally).

[79] Eggertsson (1990) calls the other models 'naïve.'

5. Voluntary transfers of property

This section examines voluntary transfers of property, including market transfers (sales) and leases (temporary transfers). It also examines laws governing inheritance, or the transfer of property from one generation to the next. For the most part, the focus is on transfers of land, though the principles are more general.

5.1. Protection of ownership and market exchange

In addition to the use and investment incentives inherent in private ownership, there are the allocation incentives inherent in the market transfer of private property rights. As the Coase Theorem implies, because transaction costs are positive and property rights are imperfect, actual market transfers must contend with various problems of enforcement. One particularly important problem in the transfer of property rights is the possibility of a claim by a previously defrauded owner. An important function of property law is to minimize this source of uncertainty over ownership, thereby facilitating market exchange (Baird and Jackson, 1984).

Information about potential prior claims on property is costly, however, so an efficient system for enforcing or maintaining ownership will balance the cost of greater certainty against the benefit. For example, consider a piece of property worth V if ownership is certain, but subject to a risk $p(x)$ that a past owner will assert a claim based on error or fraud, where x is the effort (in dollar terms) devoted to ensuring title. This might represent the cost of searching a public record of past transactions (if one exists) or of establishing verifiable proof of ownership. Assume that $p' < 0$ and $p'' > 0$. The owner's problem is to choose x to maximize $(1 - p(x))V - x$, which must satisfy $-p'(x)V = 1$, or, the marginal cost of title assurance effort must equal the increase in the expected value of the property. It follows that it is not generally optimal to eliminate all risk of loss ($p(x^*) > 0$), though owners of more valuable property will invest more to secure ownership.[80]

In terms of the law, there are two basic rules for establishing ownership of property that is sold by someone who turns out not to be the owner.[81] Under the *bona fide purchaser rule*, the buyer retains ownership of stolen property if he bought it believing the seller was the true owner, whereas under the *original owner rule*, the true owner can reclaim property by presenting adequate proof that it was wrongfully taken from him. The choice between these rules involves a trade-off. Whereas the bona fide purchaser rule inadequately deters theft, the original owner rule increases the buyers' cost of verifying that sellers have good title prior to purchase. For most property, however, the choice is not important because the likelihood that goods are stolen is small. But as property becomes more valuable, the problem of uncertain ownership becomes more important.

[80] This comes from the comparative statics derivative $\partial x^*/\partial V = -p'/p''V > 0$.

[81] This discussion is based on Shavell (2004, pp. 52–55). Also see Medina (2003).

This is especially true of land, for which an extensive public registry of ownership is maintained.[82]

5.1.1. Title systems for land

Land title in the U.S. is primarily protected by a public recording system that allows potential buyers to verify title by searching the record of past transfers, theoretically back to the root of ownership.[83] Title search is a costly process, however, especially as one goes back in time and the quality of records deteriorates. Most states therefore have enacted statutes of limitation (so-called Marketable Title Acts), or less formal guidelines (established by local bars or title insurers), aimed at limiting title searches to a reasonable length. Baker et al. (2002) develop a sequential search model (essentially a dynamic version of the model in the previous section) to characterize how far back in time buyers should search a title. They show that it is not generally optimal to search the entire record, implying that optimal search involves some residual uncertainty. A test of the model using cross-state data shows that title search guidelines vary according to the predictions of the theory. Specifically, prescribed search lengths are increasing in the cost of a title defect (as measured by title insurance premiums), the likelihood of errors in the record (proxied by the percent of developed land and the frequency of land transfer), and decreasing in title search costs.

Although the recording system is the predominant land title system in the U.S., other common law countries (and some states) have also used a system of land registration, called the Torrens system.[84] Under land registration, the government certifies ownership at the time of a transfer, thereby protecting the owner against nearly all claims; claimants can at most seek monetary compensation from a public fund (financed by registration fees). This is in contrast to the recording system, which awards successful claimants an interest in the land itself. Thus under recording landowners ordinarily purchase title insurance to provide them with financial compensation in the event of a loss.[85] More recently, Arruñada (2003) develops a model of *in rem* property rights to explain the predominance and structure of title institutions across many developed countries. Focusing on the costs and benefits of *in rem* property rights (versus contract rights), Arruñada finds that countries with recording systems tend to have a larger number of allowable property regimes compared to countries with registration systems.

Note that the two title systems provide opposing answers to the fundamental question of whom a title system should protect, the current possessor or the last rightful owner

[82] Public registries of other valuable property, like boats, automobiles, and aircraft, are also maintained.

[83] According to Black's Law Dictionary 'title is the means whereby the owner of land has the just possession of his property.' A 'deed' is a legal document which constitutes evidence of title.

[84] This system was originally instituted in Australia in 1858; see, for example, Bostick (1987) and Shick and Plotkin (1978).

[85] Miceli et al. (2002) report that in 1989 insurance claims amounted to roughly 0.05% of the total value of residential real estate transactions.

(Baird and Jackson, 1984). (In this sense, they reflect the difference between the bona fide purchaser and original owner rules.) The question is whether one is preferred on efficiency grounds. If transaction costs are zero, the Coase Theorem implies that land will be used efficiently under both systems (Miceli and Sirmans, 1995a). But since the actual transaction costs of land transfer are significant, the preferred system is the one that minimizes these costs, thereby facilitating exchange and investment.

Proponents of land registration claim that it lowers transaction costs relative to the recording system because it dispenses with the need to search anew the entire history of a parcel with each transfer (Bostick, 1987).[86] Actual attempts to compare the costs of registration and recording in those jurisdictions in the U.S. where they co-exist, however, have yielded mixed results, primarily owing to the high administrative costs of registration (Janczyk, 1977; Shick and Plotkin, 1978). Such comparisons, however, may miss the chief advantage of registration—namely, that it clears title to land in cases where land records are poor or have been destroyed. For example, land registration was instituted in Cook County, Illinois following the Great Chicago Fire, which destroyed nearly all land records. A recent study of land transactions in that county used the co-existence of both systems throughout most of the twentieth century as a natural experiment to compare land values under each system (Miceli et al., 2002). Because theory predicts that owners of higher risk properties should prefer the Torrens system, however, the empirical analysis had to control for sample selection bias in the data. Once this was done, the study found that land values in the sample were indeed higher under the registration system as compared to the recording system. Part of this increase in land value can be attributed to the protection of a current owner's subjective valuation, which can be linked to time and specific investments (Miceli, 1997, pp. 128–129). Holmes eloquently makes this point: '[M]an, like a tree in the cleft of a rock, gradually shapes roots to its surroundings, and when the roots have grown to a certain size, can't be displaced without cutting at his life (cited in Merrill, 1985b, p. 1131).'

A related issue that has been ignored in both the economics and legal literature is the method by which land is measured and transacted. In the United States there are two distinct regimes—the metes and bounds systems and the rectangular survey.[87] Under metes and bounds, property boundaries follow the natural contours of the lands and are described in terms of local, idiosyncratic features (e.g., trees, streams, rocks). Metes and bounds has been used throughout the world and remains the dominant method of defining property rights to land. The metes and bounds system has been called 'unsystematic' or 'indiscriminant' because the land is not surveyed before settlement and because the surveys are not governed by a standardized method of measurement or shape. In the United States, metes and bounds is primarily used in those states where land was granted by states rather than by the federal government (i.e., the thirteen original states as well as parts of several other states). The rectangular survey was introduced

[86] Following Barzel (1982) the recording system can be said to reduce the costs of excess measurement.

[87] See Libecap and Lueck (2006) for a history of these systems and their economic implications.

into the U.S. with the Land Ordinance of 1785, beginning in Ohio and proceeding west. The rectangular survey system uses a surveyed grid to describe land. The survey begins with the establishment of an initial point denoted by a specific latitude and longitude. From this point the land is divided into square (six miles by six miles) units called townships. Each township is divided into 36 sections; each section is one mile square and contains 640 acres. The resulting property grid is apparent in flying across the western two-thirds of the United States. Although no studies of these two methods have been completed some implications emerge. The rectangular survey is likely to lead to more market transactions, fewer border disputes and greater investment in land because the system is more well defined (e.g., the borders to not change as rivers move and trees die, there are no unclaimed parcels) and because it is a standardized rather than local system of land demarcation. Preliminary evidence from Libecap and Lueck (2006) support these implications but the topic is open for further study.

5.1.2. Title systems and development

Economists have recently begun to examine the role of land title systems in promoting economic development. For example, De Soto (2000) argues that the absence of a well-functioning system for protecting land ownership (i.e., legal title) is the single largest impediment to economic growth in most developing countries. Lack of secure title inhibits land sales, discourages investment, and prevents owners from converting land assets (which are abundant) into financial capital. De Soto's evidence is largely anecdotal, but several empirical studies have established a link between formal land title and economic investment in various developing countries (Besley, 1995; Alston, Libecap, and Schneider, 1996; Miceli, Sirmans, and Kieyah, 2001). (Also see the discussion in section 2.2.) De Soto makes the argument that legally enforced property rights are superior to those enforced by extra-legal means, thus emphasizing the economic importance of law.

5.2. Leases

A lease is a voluntary transfer of possessory rights in property (the right of use) for a limited period of time. Such an arrangement can enhance efficiency by allowing gains from specialization. The division of ownership and use, however, creates potential incentive problems for both landlords and tenants regarding the optimal maintenance and use of the property. The problem is one of moral hazard, but it is also referred to as the 'rental externality' (Henderson and Ioannides, 1983).

To illustrate, suppose that the value of a piece of property, $V(x, y)$, is an increasing function of inputs by both the tenant (x) and the landlord (y).[88] Further, suppose that V

[88] Thus, both inputs can be interpreted as maintenance. The analysis would not change if the tenant input is interpreted as the rate of utilization, which would have a negative impact on V.

is the sum of the value of the property to the tenant during the term of the lease, *T(x,y)*, and the landlord's residual value (the value of the reversion), $R(x, y)$. The first-best choices of x and y maximize the joint value of the property, $V(x, y) - x - y$, but both the landlord and tenant will make their choices to maximize their individual returns. Specifically, the tenant will choose x to maximize $T(x, y) - x - r$, where r is the rent, while the landlord will choose y to maximize $R(x, y) - y + r$. Given a fixed rent, both parties will therefore under invest in maintenance. This is because the standard fixed rent contract does not specify and enforce the first-best investment levels.[89] We will see that several aspects of lease law can be interpreted as responses to this problem.

5.2.1. *The lease: a contract or a conveyance*

Historically, all leases fell under the law of property, which viewed the lease as a conveyance of an interest in land to the tenant (Dukeminier and Krier, 2002). This gave the tenant the right to exclude the landlord from entry during the term of the lease in return for a promise to pay rent. Yet even if the tenant defaulted on the rent, the landlord could not evict the tenant; he could only sue for recovery of the rent. At the same time, the landlord had no duty to maintain the premises during the lease period. The lease thus provided strong protection of the tenant's possessory interest in the property.

In contrast, modern leases for housing are generally viewed by courts as contracts rather than conveyances.[90] This change has altered the obligations of the parties in important ways. First, landlords have a duty to maintain the property in a habitable state according to an 'implied warranty of habitability,'[91] which tenants can enforce by withholding rent. Symmetrically, however, landlords who meet their duty of maintenance can evict tenants who fail to pay rent. The obligations of the landlord and tenant, like those of the parties to a contract, are therefore mutual.

This change in the law has an economic rationale (Miceli, Sirmans, and Turnbull, 2001). Historical leases were primarily for agricultural land, and landlord inputs were relatively less important. (In terms of the above model, y did not substantially affect T.) In this context, legal protection of a strong possessory interest promoted efficient tenant investment during the term of the lease. For example, the law prohibited landlords from re-taking possession of the land after the crops were planted but before harvest. Further, tenant use ordinarily did not have a detrimental effect on the value of the reversion (i.e., x did not have a large effect on R).

The situation is different in modern real estate leases, which are primarily for housing. Here, landlord maintenance during the term of the lease is crucial, so the law has

[89] Essentially, each party to the contract faces only a portion of the marginal product of investment (T_x/V_x for the tenant and R_y/V_y for the landlord). As noted in section 5.2.3, however, a zero transaction cost contract (Cheung, 1969) would specify and enforce optimal investment levels for both parties.

[90] That is, the contract and property doctrines are merging. As Dukeminier and Krier (2002, p. 457) state: 'Is a lease a conveyance or a contract? Actually, of course, it is both.' That such a merger has occurred for commercial leases, however, is less clear.

[91] The key case is *Javins v. First National Realty Corp.*, 428 F.2d 1071 (1970). Also see Rabin (1984).

provided tenants with an enforcement mechanism by transforming the lease into a contract with an implied warranty of habitability.[92] In addition, tenant inputs are much more likely to have an effect on the value of the landlord's reversion (though in modern agriculture with sophisticated technology that can impact land this might be less true). For example, overutilization of rental housing will accelerate the rate of depreciation. The law addresses this problem with the doctrine of waste (Posner, 2003, p. 73), under which a tenant has a duty to invest in reasonable maintenance of the property. In terms of the above model, this forces the tenant to internalize the effect of his actions on the value of the reversion. The doctrine of waste and the warranty of habitability thus work in combination to create efficient bilateral incentives for maintenance in the presence of the rental externality.[93]

5.2.2. The duty to mitigate damages

Another effect of the transformation of the lease from a conveyance to a contract concerns the duty to mitigate damages. Under the law of property, landlords had no duty to mitigate damages. If the tenant abandoned the property, the landlord had no obligation to attempt to re-let it; he could just sit tight and sue the tenant for the entire rent. The transformation of the lease to a form of contract, however, imposed on landlords the contractual duty to mitigate damages by taking all reasonable steps to re-let the property.[94] The law enforces this duty by limiting the damages from tenant breach to the difference between the contract rent and the best rent the landlord could have obtained by reasonable efforts.

Mitigation of damages provides a clear economic benefit by preventing the property from being left idle. However, this raises the question of why the traditional law of leases did not impose such a duty. Economic theory suggests three possible reasons. First, agricultural tenants may have been in a better position than landlords to find substitute tenants, whereas the situation is reversed for modern residential leases. The change in the law thus simply reflects an application of the least-cost-avoider principle. Second, a duty to mitigate damages may result in inefficient re-letting of the property by landlords who mistakenly interpret tenant absence as a sign of breach. The no-mitigate rule therefore protects the tenant's possessory rights in settings where absentee use may be valuable—a situation that is more reflective of agricultural as compared to residential leases. A third possibility is that in agricultural settings the law is often less important

[92] See Hirsch (1999, Ch. 3) for an empirical analysis of the impact of habitability laws. Agricultural land leases also have implied covenants of 'good husbandry' which require the renter to maintain the quality of the land (Allen and Lueck, 2003).

[93] In this sense, the two doctrines resemble the tort rule of negligence with a contributory negligence defense, which establishes efficient bilateral incentives in accident settings. See Shavell (2004, Chapter 8) and Landes and Posner (1987).

[94] See generally Goetz and Scott (1983).

than market enforcement via repeated interaction (Allen and Lueck, 2003). For agriculture the law simply may not have developed to address this issue.[95]

5.2.3. *Cropshare versus cash rent leases in agriculture*

The choice between a cash rent lease and a cropshare lease has been an important topic since the beginning of economics.[96] Adam Smith argued that the cropshare acted as an inefficient tax on effort. Writing roughly a century later than Smith, John Stuart Mill noted that cropshare leases had an ancient origin and that the level of cultivation was not suffering. Thus, he was reluctant to claim widespread inefficiency. Smith's tax analogy, however, influenced Alfred Marshall and other neoclassical economists who later analyzed the problem. Not until Cheung (1969) extended the Coase Theorem into share cropping did the modern analysis begin. Cheung demonstrates that if transaction costs are zero, then all land leases must be equivalent, and that, therefore, the (lease) contract choice must depend on transaction costs.[97]

We present a model from Allen and Lueck (2003) that recognizes the complexity of assets and property rights to those assets as discussed in section 2.4. In both a cash rent and cropshare lease, property rights to the land are imperfect. Typically a lease agreement can only specify and enforce such basic parameters as acreage of the plot and type of crop. Such important features as soil moisture and soil nutrients cannot be economically enforced in the lease, so these attributes are essentially open access goods. In a cash rent lease the farmer pays a fixed annual amount per acre of land and owns the entire crop. As a result he supplies the optimal amount of his own inputs but overuses any inputs provided by the landowner, including the un-priced attributes of the land. In a cropshare lease, in contrast, the farmer does not pay any fee for use of the land but simply pays a predetermined share of the crop to the landowner at the time of harvest. In this arrangement the farmer and the landowner have shared ownership of the crop, so the farmer has an incentive, as Adam Smith noted, to under-provide these inputs. The farmer will also have less incentive to use inputs provided by the landowner, compared to a cash rent lease.

Consider a tract of farmland that can be used to produce crops according to $Q = h(e, l) + \theta$ where Q is the harvested crop, e is the farmer's composite input called effort, l is a composite input of land quality attributes, and $\theta \sim (0, \sigma^2)$ is a randomly distributed composite input that includes weather and pests. We assume that $h_e > 0$, $h_l > 0$, $h_{ee} < 0$, $h_{ll} < 0$, and $h_{el} = 0$, where the subscripts denote partial derivatives. The opportunity cost of the farmer's input is the competitive wage rate w per unit of farmer's effort, and the opportunity cost of the unpriced land input l is r per unit.

[95] Again, the duty to mitigate has not been universally applied to commercial leases. The reason may be that commercial leases are more like agricultural leases in terms of the factors noted in the text.

[96] Allen and Lueck (2003, chapter 4) give a detailed history of this literature.

[97] Cheung also postulated a risk-sharing effect that is discussed below.

With risk-neutral landowners and farmers, the expected profit from the farming operation is maximized, resulting in the employment of e^* and l^* units of farmer and landowner inputs. These first-best input levels are identical for the cropshare and cash rent leases and satisfy the standard conditions that marginal products equal marginal costs for both inputs. When transaction costs are positive and lease enforcement is costly, however, the input choices will be second-best. In either lease, farmers have an incentive to exploit the land's un-priced attribute because they do not face the full costs. In addition, farmers have an incentive to under-report the output in the cropshare lease.

For the cash rent lease, the farmer owns the entire crop and chooses his inputs to maximize expected profit. Because the farmer does not have indefinite tenure of the land, he does not face the true opportunity cost of using the attributes of the land. If we denote the reduced costs he faces as $r' < r$, the farmer's objective is:

$$\max_{e,l} \Pi^r = h(e,l) - we - r'l. \tag{5.1}$$

The second-best solutions e^r and l^r satisfy $h_e(e) = w$ and $h_l(l) = r'$. Since $h_{el} = 0$, we note that the farmer's input level is identical to the first-best optimum; that is, $e^r = e^*$. However, since $r' < r$, the land is over-worked ($l^r > l^*$) because the farmer does not face the full cost of using the land's attributes (i.e., he ignores the value of the reversion).

In a cropshare lease, the farmer receive sQ and the landowner receives $(1 - s)Q$, where $0 < s < 1$. The farmer's objective is:

$$\max_{e,l} \Pi^s = s[h(e,l)] - we - r'l. \tag{5.2}$$

Now the second-best solutions e^s and l^s satisfy $sh_e(e) = w$ and $sh_l(l) = r'$. These solutions indicate that the farmer supplies too few of his inputs because he must share the output with the landowner; that is $e^s < e^*$. As with cash rent, the farmer over uses the land attributes, or $l^s > l^*$; however, since $l^r > l^s > l^*$, the use of the land is less excessive than it is with cash rent. This means that although a share lease still provides the farmer with an incentive to over use the land, this incentive is not as powerful as it is with the cash rent lease.

Farmers and landowners choose the lease that maximizes the joint expected return to the tract of land. This requires comparing the expected net return to the land in both leases, where the net return is given by the appropriate indirect objective function. For the cash rent lease,

$$V^r(w,r,r') = h(e^r,l^r) - we^r - rl^r. \tag{5.3}$$

With the cropshare lease there are additional costs of measuring and dividing the harvested crop (Barzel, 1982; Homstrom and Milgrom, 1991). These costs are given by μ so that the net value function is,

$$V^s(w,r,r',\mu) = h(e^s,l^s) - we^s - rl^s - \mu. \tag{5.4}$$

The joint maximization problem is $\max\{V^r, V^s\}$. The tradeoff between the two leases is straightforward.[98] The benefit of cash rent is the avoidance of the costs of dividing the harvested output. The benefit of cropshare is the reduction in the total distortion of input levels. Thus cropsharing should be observed when output measurement costs are low, and when soil attributes are easy to exploit. Cash rent leases should be observed under the opposite conditions. The effect of parameter changes on the net value of each contract can illuminate this tradeoff and lead to hypotheses about lease choice.

Consider first how changes in μ affect V^r and V^s. The net value of the cash rent lease V^r does not depend on output division costs. The net value of the crop share lease V^s however, declines as these costs increase. By the Envelope Theorem $\partial V^s/\partial\mu < 0$. This implies that as the costs of output division increase it is less likely that the cropshare contract will be chosen. The comparative statics for r are similar. By the Envelope Theorem $\partial V^s/\partial r = -l^s$ and $\partial V^s/\partial r = -l^r$. Because neither l^s nor l^r depend on r, the second derivatives of V^s and V^r with respect to r are zero. Therefore, V^s and V^r are linear functions of r. Thus, an increase in the cost of land attributes will lower the value of either lease (holding r' constant), but it will lower the value of the cash rent lease more because land inputs are used more intensively in a cash rent lease than in a cropshare lease ($l^r > l^s$). This implies that a cropshare lease is more likely to be chosen both as the unpriced attributes of the land become more easily damaged, and as land value increases.

Allen and Lueck (2003) find support for these predictions using data from North America and evidence from around the world. They show that cropshare leases are more likely when crop division costs are low and where the ability of farmers to adversely affect the soil is high. Further, cash rent leases often contain clauses that discourage exploitation of the soil. For example, hay crops are more susceptible to under-reporting since they are used on the premises, and thus are more often cash-rented. Land used for row crops is more susceptible to overuse than is land used for grains, and the data show that row crops are more likely to be cropshared.

The property rights–transaction cost approach to leases assumes that everyone is risk neutral, and relies on a trade-off between different incentive margins to explain lease terms. This approach contrasts with the more common approach—the traditional Principal-Agent (P-A) model—which assumes leases (or contracts in general) are designed to balance risk sharing against moral hazard. Despite the prominence of the risk-sharing paradigm (e.g., Newberry and Stiglitz, 1979; Hayami and Otsuka, 1993), the empirical evidence to support its implications is scarce, especially for agriculture. In one of the early studies to confront risk-sharing and contract choice, Rao (1971) found that crops with high yield and profit variability were less likely to be sharecropped than crops with low yield and profit variability—a refutation of the P-A model. Using data from several thousand farmland leases, Allen and Lueck (1999, 2003) present a series of empirical tests that find virtually no support for the risk-sharing approach. In a variety

[98] The formal comparative statics predictions are derived in Allen and Lueck (2003, chapter 4).

of empirical tests, Allen and Lueck find no support for the general hypothesis that share leases are more likely to be chosen over cash rent leases when crop riskiness increases. In fact, there is evidence that the relationship is the opposite; that is, as crop riskiness (in terms of yield variability) increases, cash rent leases are often more likely (Allen and Lueck, 1995, 2003, and Prendergast 2000, 2002). This result holds across all crops and regions examined in Allen and Lueck (2003).[99]

Compared to the basic P-A model, the transaction cost approach does not explicitly distinguish between principals and agents, nor does it make differential assumptions about the risk preferences of the contracting parties. In modern farming it is especially difficult to establish such a dichotomy because farmers and landowners have nearly identical demographic characteristics. Both farmers and landowners make decisions, so formal models more in line with double moral hazard are more appropriate (e.g., Eswaran and Kotwol, 1985; Prendergast, 2002). More importantly, by diverting attention away from risk-sharing—which is hard to test and has thus far generated little empirical support—the approach opens the door to a wider array of pure incentive effects that shape organization.

5.3. Inheritance of land

Inheritance rules govern the intergenerational transfer of land and other property. They thus represent an important means of voluntary transfer of property. As such, one function of these rules is to ensure that the wishes of testators (i.e., current owners) regarding the disposal of their property are fulfilled. An offsetting concern is to limit the extent to which the 'dead hand' can constrain the uses of property into the uncertain future (Stake, 1990, 1998a). In attempting to balance these goals, Anglo-American law gives testators considerable freedom in the disposal of their property, but imposes some constraints, including primogeniture and the Rule Against Perpetuities, which we discuss here.

The rule of primogeniture, under which all property passes to a decedent's eldest son, was the predominant rule in early English common law and has also been used in cultures throughout the world.[100] The most common economic explanation for the rule is that it prevents inefficient fragmentation of land (Posner, 2003, p. 517). There are, however, two objections to this rationale. First, a well-functioning land market should allow entrepreneurs to counteract the effects of fragmentation. Thus, we would expect the rule to be most prevalent in societies where land markets are primitive or do not exist. (Baker and Miceli (2005) provide evidence for this prediction.) Second, even if scale

[99] Outside the area of agriculture a series of papers have found similar results (see the summary in Prendergast, 2002). Ackerberg and Botticini (2002), however, show that risk sharing might still be important in contract choice if one takes into account the endogenous matching of farmer with different risk preferences and land suitable to crops of varying risk. Nearly all of this literature can be criticized though for data that does not reliably measure exogenous risk.

[100] Alston and Schapiro (1984) study the demise of primogeniture in the United States.

economies are important, why constrain a testator's choice of the most suitable inheritor? In particular, why not adopt a 'best-qualified' rule that maintains scale economies while expanding the testator's options? One possible explanation is that such a rule might promote wasteful rent seeking by competing heirs (Buchanan, 1983).

Another constraint on a testator's discretion is the Rule Against Perpetuities, which limits restrictions that can be imposed in a will to a set period of time, equal to the lifetime of anyone alive when the will was created plus twenty-one years.[101] This time limit reflects offsetting economic factors (Shavell, 2004, pp. 67–72).[102] On one hand, testators (current owners who die with a will) should have broad control over the disposition of their property after their death, both to maximize their utility and to give them an incentive to acquire and create wealth during their lifetimes. Such control is especially beneficial if immediate heirs are known to be spendthrifts. On the other hand, testators may not be able to foresee the best uses of their property in the uncertain future, or may specify uses that future generations will deem harmful (e.g., imposing conditions for use based on race or religion), or create constraints that are extremely costly to undo.[103] A final reason to limit testators' discretion is simply to preserve some amount of intergenerational equity in the distribution of wealth, given that the current generation, by definition, controls all wealth.

6. Involuntary transfers of property

This section examines involuntary transfers of property from one private party to another. (We examine involuntary transfers from private parties to the state (takings) in section 8.) Initially, we discuss transfers that occur as a result of uncertainty about ownership or boundary location, and hence, for the most part, are unintentional. We conclude by discussing intentional involuntary transfers, or theft.

6.1. Adverse possession

Adverse possession is a curious doctrine that appears to legitimize the theft of land by squatters. The doctrine establishes title in property to the current user or possessor without the consent of, or compensation to, the original legal owner. It therefore has little rationale in the absence of transaction costs and is viewed typically as a method of clarifying title that has become clouded over time (Dukeminier and Krier, 2002). In order to gain title the adverse possessor must 'openly and notoriously' maintain exclusive possession for a statutorily specified term that ranges from one to thirty years in

[101] Black's defines the rule as the '[p]rinciple that no interest in property is good unless it must vest, if at all, not later than 21 years, plus a period of gestation, after some life or lives in being at the time of the creation.'

[102] Also see Ellickson (1986), Epstein (1986), and Dukeminier and Krier (2002).

[103] The doctrine of *cy pres* allows courts to substitute related, but less harmful, uses of bequeathed property in situations where prescribed uses are offensive to the current generation.

the United States. The precept of adverse possession is embedded in the common law and can be traced to an English statute enacted in 1275. Contemporary American law is a mixture of statutory and case law in which statutes define required time periods and other specific conditions, while court decisions define 'notorious' possession and other less specific requirements.

As discussed above, adverse possession is recognizable as a first possession doctrine. The adverse possessor has 'relative title,' by virtue of prior possession, or has 'rights against the rest of the world from the moment that he claims possession' (Epstein, 1986, p. 675.) Moreover, in a successful adverse possession action the original owner's title is deemed to be invalid. Consequently, first possession becomes an accurate description of the process by which ownership is established. The law essentially treats the property as abandoned by the original owner. Historical adverse possession cases have dealt with such issues as abandoned farmland, cabins in the woods, and old mining sites. Typical cases today deal with title to real estate in situations where property boundaries are either unknown or misunderstood. For example, a homeowner builds an addition that, it turns out, is actually on the neighbor's legal property. Under adverse possession the homeowner gains title to the property in question by virtue of his possession, through the addition, for the duration of the statutory period. In the historical cases, heterogeneity probably served to mitigate dissipation from first possession, and there is little evidence of racing among potential adverse possessors. In the modern real estate boundary cases, heterogeneity is at its extreme. There is only one potential claimant; hence, there is no dissipation.

Several theories have arisen to explain the details of adverse possession doctrine, treating it as a time-limited property right.[104] The most common are that it lowers the transaction costs of clearing title to land, and that it prevents valuable land from being left idle. These arguments are not entirely convincing, however, given the quality of land records in most jurisdictions and the option value of leaving land undeveloped.

A more convincing argument is based on the presence of offsetting risks to ownership of land. The first risk arises from the possibility, discussed in section 5, of past claims by previous owners who were deprived of their title through fraud or error. A time limit on such claims limits this risk to current owners. Specifically, let $p(t)$ be the risk of such a claim, where t is the duration of the prior owner's property right as specified by the adverse possession statute. We assume that $p'(t) > 0$, reflecting a higher risk to the current owner for longer-lasting property rights, and $p(0) = 0$.[105] The other risk is that the current owner may himself be displaced by a squatter. This possibility can be reduced, however, by periodic monitoring of the property to eject squatters or correct boundary errors (Ellickson, 1986). A longer time limit on the current owner's property right lowers this cost by reducing the required frequency of monitoring. Formally, let

[104] See, for example, Ellickson (1986), Merrill (1985b), Miceli and Sirmans (1995b), and Stake (2001).

[105] Note that land registration under the Torrens system effectively sets $t = 0$ by extinguishing most past claims.

$m(t)$ be the cost of monitoring that the current owner must spend to retain title with certainty, where $m' < 0$ and $m(\infty) = 0$.

Now suppose the current owner contemplates investing in the land. Let $V(x)$ be the market value of an investment of x dollars, where $V' > 0$ and $V'' < 0$. Given uncertainty, the owner will choose x to maximize the expected value of the property, $(1 - p(t))[V(x) - m(t)] - x$, taking t as given. This yields the first-order condition

$$(1 - p(t))V'(x) - 1 = 0. \tag{6.1}$$

Condition (6.1) defines the optimal investment, $x^*(t)$, as a function of the time limit, where $\partial x^*/\partial t = p'V'/(1 - p)V'' < 0$. Thus, increasing the duration of property rights actually *reduces* investment incentives by increasing the risk of a past claim.

Given this characterization of the landowner's problem, we can derive the optimal duration of property rights as the value of t that maximizes the total value of the land net of monitoring costs:[106]

$$V(x^*(t)) - x^*(t) - m(t). \tag{6.2}$$

Differentiating (6.2) and substituting from (6.1) yields

$$p(t)V'(x)(\partial x^*/\partial t) = m'(t). \tag{6.3}$$

Thus, the optimal time limit balances the detrimental effect of longer t on investment incentives (the left-hand side) against the savings in monitoring costs (the right-hand side).

Although all fifty states have adverse possession statutes, as noted, the length of the statutory period varies, ranging from one to thirty years with mean length of 13.63 years.[107] Two empirical studies of adverse possession statutes show that this cross-state variation is broadly explained by the economic model (Netter, Hersch, and Manson, 1986; Baker et al., 2001). In particular, states with slower urban growth rates and lower per acre farm values (reflecting a lower value of development) have longer statute lengths, while states with more efficient legal systems (suggesting lower monitoring costs) have shorter statute lengths.

6.2. *The mistaken improver problem*

The analysis to this point has treated the probability of a competing ownership claim as a function only of the statutory period, but owners can actively lower the risk of a claim by surveying the property prior to development to detect boundary errors and eject squatters, or by searching the land records (as discussed in section 5) to uncover title errors. Suppose that a survey reveals ownership with certainty. If the developer

[106] We assume that whoever ends up as owner will spend $m(t)$.

[107] The data are from Leiter (1999). In some states, the length is conditional on whether the squatter has "color of title" (i.e., evidence that appears to, but does not legally, convey title).

is the owner, he can proceed with development as if there is no risk of a loss,[108] whereas if the survey reveals someone else to be the owner, the developer can purchase the land if it is more valuable in a developed state. In this way, the value of the land is maximized. Determining ownership is costly, however, which may make it more profitable for the developer to proceed without a survey. This raises the possibility of mistaken improvement of another's property—the so-called mistaken improver problem.

To examine this problem formally, let V be the market value of a parcel of land if it were to be improved, and let p be the probability that the land is owned by someone else who values it in its currently unimproved state at R. Further, suppose R is unobservable to the developer but is known to vary according to the distribution function $F(R)$. If the developer surveys at cost s prior to developing (in which case he learns R and only develops if $V > R$), the expected value of the land is $(1-p)V + p\mathrm{E}\max[V, R] - s$, or

$$(1-p)V + p\left[F(V)V + \int_V^\infty R\,dF(R)\right] - s. \tag{6.4}$$

Expression (6.4) has three parts: the value if the developer is the true owner, the value if someone else is the owner, and survey costs. If, however, the developer proceeds without a survey, the value of the land is fixed at V, regardless of who turns out to be the owner (given the irreversibility of development). A survey is therefore optimal if (6.4) exceeds V, or if

$$p\int_V^\infty (R-V)\,dF(R) > s. \tag{6.5}$$

The left-hand side of this condition is the expected benefit of avoiding irreversible improvement of the land when it is owned by someone else who values it more highly in its unimproved state.

Developers will not necessarily make the first-best survey decision on their own, however, because they will ignore the opportunity cost of development when someone else is the owner. The law, however, provides victims of mistaken improvement remedies that potentially create the right incentives. The law of mistaken improvement dates back at least to Roman times, where the law of accession stated that materials affixed to land became the property of the owner. The mistaken improver could at most seek compensation for the value of the improvements. The modern law in most states is dictated by so-called betterment acts, which typically allow landowners the option of either paying for the improvements (according to the old rule), or forcing the improver to buy the land at its unimproved value (Dickinson, 1985).[109] It turns out that this 'option' remedy induces would-be improvers to internalize the opportunity cost of the improvements in

[108] In that case, he will invest an amount $x^* > x^*(t)$ for any $t > 0$.

[109] Dukeminier and Krier (2002, pp. 152–53) also note that common law was originally 'harsh' in that the improver always lost the land and the improvement, but modern cases grant relief to the 'innocent improver.'

the face of ownership uncertainty and hence gives them exactly the right incentives to conduct a survey (Miceli and Sirmans, 1999).

6.3. Partition of real estate

Another form of involuntary transfer, this time involving joint owners of property, is the right to partition real estate. Under the common law, each co-owner of a parcel of land has the right to force a physical partition of the property (partition in kind) into separately owned parcels. While this solution overcomes transaction costs among co-owners (due, for example, to the anti-commons problem discussed in section 2.2), it may result in excessive fragmentation if there are scale economies associated with the best use of the land. State partition statutes have sought to address this problem by providing courts with an alternative to in-kind partition–namely, forced sale of the undivided parcel with division of the proceeds to the co-owners in proportion to their ownership shares.

The problem with forced sales, however, is that non-consenting owners only receive the market value of their shares, thus depriving them of any subjective value that they may attach to the land.[110] In terms of efficiency, forced sale will only be preferred to partition in kind if the preserved scale economies exceed the foregone subjective value of all non-consenting owners (Miceli and Sirmans, 2000). Courts seem sensitive to this trade-off. In particular, they tend to favor partition in kind, unless the resulting fragmentation would *materially* reduce the aggregate value of the land.[111] This standard offers courts a margin for protecting subjective value of non-consenting owners against expropriation.

6.4. Theft

The most obvious form of involuntary transfer of property is theft, which is classified as a crime. The economic theory of criminal enforcement is well developed (see Shavell, 2004, Chapters 20–24). Here we comment on the intersection of criminal law and property law. In economic terms, the transfer of property by theft presents the following paradox—if a thief values the stolen property more than the owner does, then the transfer is efficient (though coercive). Thus, why not simply force the thief to pay a fine equal to the value of the stolen property, in effect, treating the theft as a tort in which the state is not involved in enforcement and policing? One objection is that the thief will sometimes avoid detection, thus lowering his expected cost and allowing some inefficient transfers, but this problem could be addressed by simply inflating the fine in proportion to the inverse of the probability of detection.[112]

[110] In effect, forced sales substitute liability rule protection of owners' shares for property rule protection, thus creating the possibility of an inefficient sale (Calabresi and Melamed, 1972).

[111] See, e.g., *Trowbridge v. Donner*, 40 N.W.2d 655 (1950).

[112] The economic theory of punitive damages has a similar rationale.

A more fundamental objection to the 'efficient theft' argument is that it permits individuals to substitute coercive transfers for market transfers (Calabresi and Melamed, 1972; Klevorick, 1985; Coleman, 1988). Market transfers are generally more efficient than coercive transfers because courts may err in setting the right amount of compensation (the standard problem with liability rules), and because owners, fearing such a transfer, will devote excessive resources to the protection of their property (a form of rent seeking).

If the preceding argument makes sense for tangible property, it is all the more persuasive when the violation concerns one's bodily integrity or civil rights. The law therefore seeks to deter such violations by setting the penalty above compensatory damages (and possibly including the risk of imprisonment) while labeling them as crimes (illegitimate transfers). (See the discussion of inalienability in section 9.)

For real property, theft is a somewhat different phenomenon, because the asset (land) cannot be moved from its current location. This means that for real property, 'theft' is really damage to the asset—which is often handled by nuisance law (as discussed in section 7.4), or removal of some part of the asset (e.g. fence, game, timber)—which is a more typical criminal act.

7. Land use conflicts: externalities and property

Externalities arise when one party uses his property in a way that imposes a cost (or confers a benefit) on another party without first obtaining that party's consent. (In this sense, externalities are a form of involuntary transfer.) When assets are complex and transaction costs are positive, externalities are ubiquitous. As we noted above, externalities might be viewed as 'theft' for the case of an immoveable asset. This is because property rights to at least some of the attributes of an asset will be imperfect and thus contain problems of open access or moral hazard. In the case of land, externalities are important since any parcel (except an island or continent) will have neighboring owners, but they also arise in the context of air quality, noise, and water, where property rights are especially hard to define and enforce.

In this section, we analyze various remedies for externalities (primarily harmful externalities), focusing specifically on a comparison of the standard tax-subsidy approach most commonly associated with Pigou, with the property rights, or Coasian, approach.[113] We also discuss the common law remedies of trespass and nuisance, public controls like zoning, and private responses like covenants.

7.1. A model of externalities in the short and long run

In this section we develop a model of external costs that we will use to examine the various remedies just described. We examine the general case of 'bilateral care' externalities in which both parties can affect the amount of the damage. We also consider

[113] The analysis is based on Polinsky (1980) and White and Wittman (1979).

both short and long run notions of efficiency in anticipation of the fact that some remedies that are efficient in the short run are inefficient in the long run. To be concrete, consider, as did Coase (1960), a railroad whose trains emit sparks that occasionally set fire to crops on farmland adjacent to the tracks. Suppose that the number of trains being run is n_T and the number of farms (or total acreage) is n_F, resulting in crop damage equal to $n_T n_F D(x, y)$, where D is the damage (in terms of reduced crop value per acre) each train causes, x is dollar spending on precaution per train by the railroad (e.g., whether it installs a spark arrester), and y is dollar spending on precaution by each farmer (e.g., where he locates his crops).[114] We assume that $D_x < 0$, $D_y < 0$, $D_{xx} > 0$, and $D_{yy} > 0$, reflecting diminishing marginal benefits to precaution. The benefits of railroading and farming are captured by $b_T(n_T)$ and $b_F(n_F)$, which are the marginal benefit functions for the two activities, respectively, both of which are assumed to display diminishing marginal benefits (i.e., $b'_j < 0$, $j = T, F$). The total value of the two activities is given by

$$W = \int_0^{n_T} b_T(u)du + \int_0^{n_F} b_F(z)dz - [n_T n_F D(x, y) + n_T x + n_F y]. \tag{7.1}$$

In the short run, the numbers of trains and farms are fixed.[115] Thus, the first-order conditions only concern the expenditures on precaution (x, y) that maximize (7.1) and are given by

$$n_F D_x(x, y) + 1 = 0 \tag{7.2}$$

$$n_T D_y(x, y) + 1 = 0. \tag{7.3}$$

These conditions state that the parties should invest in precaution up to the point where marginal benefits in terms of saved damages equal marginal costs. In the long run all assets become choice variables so the number of trains and farms (n_T, n_F) must also be chosen to maximize (7.1). The resulting first-order conditions for n_T and n_F are

$$b_T(n_T) - [n_F D(x, y) + x] = 0 \tag{7.4}$$

$$b_F(n_F) - [n_T D(x, y) + y] = 0, \tag{7.5}$$

which state that each activity should be increased to the point where the last unit (trains or farms) yields zero profit.

[114] This formulation of expected damages assumes constant returns to scale in number of trains and farms. See Shavell (2004) for a similar model in the context of tort law.

[115] The 'short run' is the same as economic models of torts which hold the 'activity levels' (e.g., automobile miles driven) fixed (see Shavell (2004, Chapter 8) and Landes and Posner (1987)). In tort models the 'long run' means that activity levels are endogenous.

7.2. The Pigovian tax-subsidy approach

The traditional (pre-Coase) approach to the control of externalities is the Pigovian, or tax-subsidy approach. The idea is that the government needs to impose a tax on, or pay a subsidy to, the source of the externality (the railroad in this case) in order to force it to internalize the damage that it causes. In a model in which damage depends on the actions of both parties it becomes immediately clear that 'causation' is ambiguous, as Coase (1960) first noted.

Consider first short run incentives regarding precaution, holding the number of trains and farms fixed. Under a tax, the railroad pays the government based on damages imposed. Both the railroad and farmer will choose efficient care under this remedy provided that, first, the marginal tax equals the marginal damages imposed on farmers (from (7.2), $t'(x) = n_F D_x$), and second, that farmers do not receive the revenue from the tax (except possibly as a lump sum payment). Symmetrically, a subsidy scheme under which the government pays the railroad to reduce crop damage achieves bilateral efficiency in the short run provided that the marginal reduction in the subsidy equals marginal damages (i.e., $-s'(x) = n_F D_x$).

Note that the structures of the tax-subsidy schedules are not fully determined by these conditions because the tax only impacts the marginal conditions for the choice of the inputs x and y. This is not the case, however, when we take into account long run efficiency. Consider first the railroad's decision about the number of trains. According to condition (7.4), the railroad will only choose the efficient number if it internalizes the full cost of the crop damage per train. This requires that it pay a tax per train equal to $n_F D(x, y)$. (Note that this tax satisfies the marginal condition above.) Clearly, a subsidy that involves any payments to the railroad will therefore result in too many trains. As for farming, condition (7.5) says that efficient entry of farmers requires that each farmer internalize the crop damage that his entry contributes to total damages. This condition is satisfied as long as farmers do not expect to receive any compensation for their losses (including lump sum compensation). In combination, these results show that only a tax scheme can achieve bilateral efficiency in both the short and long run.

7.3. The property rule–liability rule framework

As discussed above, one of the contributions of Coase (1960) was to challenge the Pigovian assumption that externalities necessarily lead to market failure. This recognition suggests an expanded set of remedies for controlling externalities, which is best exemplified by the choice between property rules and liability rules (Calabresi and Melamed, 1972).[116] Under property rules, right holders can refuse any unwanted infringements of their rights, enforceable by injunctions (or criminal sanctions in the case

[116] Also see Polinsky (1980) and Kaplow and Shavell (1996) for more recent analyses of property rules versus liability rules.

of theft). Property rules thus form the legal basis for voluntary (market) exchange of rights. In contrast, liability rules do not entitle right holders to refuse infringements of their rights; instead, they can only seek monetary compensation in the form of damages. Liability rules thus form the basis for court-ordered or non-consensual transactions.[117] Together both types of rules define a property system seemingly designed to allocate resources to their highest valued uses in the presence of varying transaction costs.

As Kaplow and Shavell (1996) note, when transaction costs are zero, property rules and liability rules should be equally efficient because the Coase Theorem applies. The choice thus turns on transaction costs, particularly the costs of contracting, the costs of court adjudication, and legal administration. When contracting costs are relatively low, property rules are preferred because they ensure that all transactions are mutually beneficial. When contracting costs are high (e.g., public nuisance cases), however, the costs of reaching an agreement under property rules may prevent otherwise efficient transactions from occurring.[118] Liability rules have an advantage in this case because they allow the court to force a transfer. In this way, a court-ordered transaction replaces a market transaction. This advantage of liability rules, however, needs to be weighed against the possibility of court error in setting damages, which may result in too many or too few transactions. Furthermore, because liability rules require courts to establish the initial terms of a transaction by setting damages (which the parties may later adjust), the administrative costs of using this rule will likely be higher than under a property rule (Kaplow and Shavell, 1996).

In the context of the railroad-farmer conflict, a liability rule entitles farmers (victims) to seek monetary compensation for their damages but not to stop the damage from occurring.[119] If liability is strict, the railroad (injurer) must pay full compensation regardless of its level of precaution. In terms of short run efficiency, strict liability induces efficient precaution by the railroad, but because farmers are fully compensated, they have no incentive to take precaution. (The outcome is identical to a tax scheme where the revenue is paid to victims as compensation.) In contrast, a negligence rule, which only holds the railroad liable for damages if it takes less than the efficient level of abatement as defined by (7.2) (for example, if it fails to install spark arresters), will induce both parties to take efficient care. The railroad will take care to avoid liability, and the farmers will take care to minimize their losses.[120]

Neither liability rule, however, will achieve long run efficiency. Under strict liability, too many farmers will enter because they do not consider the impact that their entry has on total damages. Although the railroad does face full liability for each train that

[117] Calabresi and Melamed (1972) also discuss a third rule, an inalienability rule, which prevents transfer of right under any circumstances (including consensual transfers). We discuss this rule in section 9.

[118] The issue here is identical to contracting problems association with such large scale resources such as air, groundwater, oil, and wildlife as discussed in section 4.4.

[119] Though as noted, the Coasian tradition would not use the term victim given the 'reciprocal nature' of the externality problem.

[120] See Shavell (2004, Chapter 8) and Landes and Posner (1987) for a fuller discussion of the various negligence rules.

it runs, equal to $n_F D(x, y)$, this amount is too large because of the excessive number of farms. Thus, too few trains will run (though the number of trains will be efficient, given the number of farms). The situation is reversed under a negligence rule. The railroad will invest in optimal abatement to avoid liability, but as a result, it will run too many trains (Polinsky, 1980b). In contrast, farmers will face the full amount of their damages, $n_T D(x, y)$, but too few farmers will enter because the number of trains is too large. In general, liability rules cannot create efficient long run incentives because of the constraint that what one party pays the other must receive.[121]

If the farmers' rights are protected by a property rule, they can block the railroad from running any trains by means of an injunction. The railroad, however, can seek to purchase rights to impose crop damage. For each train that it runs, the railroad will invest in abatement up to the point where the last dollar spent just equals aggregate marginal damages to all farmers, after which it will prefer to compensate farmers for the residual damages. Then, given efficient abatement per train, the railroad will run trains up to the point where the aggregate amount it has to compensate farmers equals the marginal benefit of one more train. This results in the efficient number of trains.[122]

Efficient precaution by farmers can similarly be achieved by contracting. This requires that the railroad compensate farmers for their costs of precaution up to the point where the last dollar spent on precaution equals the marginal reduction in aggregate damages owed. Achieving the efficient *number* of farms (or acres) is much more problematic, however. According to condition (7.5), long run efficiency requires that farmers enter (or add acreage) up to the point where the marginal benefits of the last farm equal its marginal contribution to crop damage plus cost of precaution. But since farmers are compensated for these costs under the current assignment of rights, there exists an incentive for too many to enter. In theory, private contracting can prevent excessive entry, but only if the railroad can identify all *potential* entrants into farming and offer to pay them their marginal benefit of entry if they agree to stay out. Clearly this poses a significant informational demand on the railroad. (Of course, a similar problem faces farmers if the property right is initially assigned to the railroad.) This discussion illustrates the limits of private contracting in internalizing externalities, especially regarding long run efficiency (Frech, 1979; Wittman, 1980; Holderness, 1989), though these limits must be compared to the limits of public action in determining the optimal second best remedy.

7.4. The law of trespass and nuisance

The primary common law responses to externalities are trespass and nuisance. Specific examples of trespass are squatters and boundary encroachment, while examples

[121] This is an example of the *paradox of compensation* (e.g., Cooter and Ulen, 1999, p. 169) which is also found in tort law and contract law remedies (Cooter, 1985). It can be avoided by 'de-coupling' liability and compensation, or with a contract or compensation mechanism that defines and enforces the optimal choices for both parties.

[122] Another possibility is that the railroad could buy all the land and engage in farming, or farmers could collectively buy and manage the railroad.

Table 7.2
Thresholds distinguishing trespass and nuisance

Trespass	Nuisance
Defendant's act occurs on plaintiff's land	Defendant's act occurs on defendant's land
Harm is 'direct'	Harm is 'indirect'
Invasion by 'tangible' matter	Invasion by 'intangible' matter
Interference with 'exclusive possession' of land	Interference with 'use and enjoyment' of land

of nuisance are air, water, and noise pollution. More generally, Table 7.2 lists several thresholds that the common law has developed for distinguishing between the two doctrines (Merrill, 1985a; Miceli, 1997, p. 119).

The primary remedy under trespass is an injunction against the unwanted intrusion.[123] Thus, the landowner's right to exclude is protected by a property rule. The remedy under nuisance law is more complicated. First, the landowner can only obtain relief if the invasion is substantial, and even then, he may have to be satisfied with money damages (a liability rule). If the landowner wishes the harm to be enjoined, he must meet the further legal standard of showing that the harm outweighs the benefit of the nuisance-creating activity (Keeton et al., 1984, p. 630).

Merrill (1985a, 1998) argues that this distinction between trespass and nuisance can be broadly understood in terms of the choice between property rules and liability rules. Cases of trespass ordinarily involve a small number of parties where the intruder is easily identifiable.[124] Again, as we discussed above, contracting costs among the parties tend to be low, and property rules are the preferred remedy. In contrast, cases of nuisance often involve large numbers or sources of harm that are difficult to identify. Thus, transaction costs are high and contracting is unlikely to lead to the efficient outcome. In cases like this liability rules are preferred.

The well-known case of *Boomer v. Atlantic Cement Co.* provides an illustration of this choice.[125] The case involved a group of landowners who sought an injunction against a large cement factory because of the dirt, smoke, and noise that it produced. The court denied the injunction and instead awarded money damages on the grounds that the injunction would have forced the factory to shut down, causing a loss of jobs and the company's substantial capital investment. The court's decision seems correct in view

[123] The doctrine of necessity noted in section 2.5 indicates that not all physical invasions are considered trespass.

[124] The law also distinguishes 'private' from 'public' nuisances (Dukeminier and Krier, 2002, pp. 773–774). A 'public' nuisance represents a broader notion of harm borne by the general public rather than one or a few landowners. In the cases we examine below, *Spur* represents both a public and private nuisance case, while *Boomer* is a public nuisance case. As we note in section 7.5, zoning seems to emerge in cases of public nuisances.

[125] 26 N.Y.2d 219, 309 N.Y.S.2d 312, 257 N.E.2d 870 (Court of Appeals of New York, 1970).

of the high costs of contracting among the large number of affected homeowners that would have been necessary to keep the plant operating under an injunction.

The equally famous case of *Spur Industries v. Del Webb*[126] also illustrates the important issues but with a slightly different result. In this case a cattle feedlot (Spur) northwest of Phoenix had been in operation prior to the development of homes by Del Webb. As the home development expanded toward the feedlot homeowners became increasingly impacted by the smell of cattle manure. Del Webb filed suit on the basis of a public and private nuisance. The court agreed with the litigants that there was a nuisance and that the feedlot activity should be enjoined, but in addition it forced Del Webb to indemnify Spur for the cost of moving or closing the feedlot. In this manner the court used a combination of property and liability rules. It partly invoked the 'coming to the nuisance' defense (a property rule), which states that a party with a prior activity cannot be liable for a nuisance. (In so doing, it effectively labeled the feedlot as the 'victim'.) The coming to the nuisance doctrine has a simple economic rationale in that a late-comer to an area impacted by nuisance activities will be faced with land prices that capitalize the reduction in value from the nuisance and thus later damages would be overcompensation.[127] However, in awarding damages (a liability rule) the court also recognized the costs of organizing a buyout by the many homeowners in the subdivided development, given that the feedlot had become an inefficient land use.

7.5. Zoning, covenants, and common law control

Probably the most common legal response to urban land market externalities in the United States is zoning, which is a form of public regulation.[128] The economic rationale for zoning is that 'similar land uses have no (or only small) external effects on each other whereas dissimilar land uses may have large effects' (White, 1975, p. 32), creating what the common law calls a 'public nuisance.' The widespread use of zoning, however, does not necessarily make it the most efficient response to externalities. High administrative and enforcement costs often exceed the saved 'nuisance costs' (Ellickson, 1973). This would not be a problem, however, if the penalty for violations were payment of an appropriate fine, which would allow landowners to circumvent inefficient regulations. In this sense, zoning regulations are best enforced by a liability rule (White and Wittman, 1979). The fact that compliance with zoning ordinances is required, however, (that is, they are enforced by a property rule) forecloses this route to efficiency.

[126] 108 Ariz. 178, 494 P.2d 700 (Supreme Court of Arizona, 1972).

[127] See Wittman (1980) and Pitchford and Snyder (2003) for economic models of this doctrine. Pitchford and Snyder (2003) show how the 'coming to the nuisance' doctrine can lead to overinvestment by the first party in a sequential model of land use, and argue that the ruling in *Spur* fits their framework.

[128] Zoning was declared constitutional in *Village of Euclid v. Ambler Realty*, 272 U.S. 365 (1926). Economic analyses of zoning with a focus on property rights include Fischel (1985) and Nelson (1977). Siegan (1972) studies Houston, the only large American city without municipal zoning and finds that the lack of zoning has not adversely impacted land use.

A private alternative to zoning is the use of land use servitudes (e.g., covenants, easements, or equitable servitudes) that impose limits on what landowners can do with their property.[129] Such restrictions are frequently observed in condominiums, coops, homeowner associations, and other common interest communities, which comprise a growing portion of the housing market (Dwyer and Menell, 1998: 808–887; Hansmann, 1991). The economic function of these restrictions is twofold: to overcome free rider problems in the provision of certain jointly consumed amenities (De Geest, 1992); and to internalize neighborhood and rental externalities (Cannaday, 1994; Hughes and Turnbull, 1996). Since these covenants overcome 'market failures' associated with ordinary housing markets, developers can charge a premium for them. Further, since the restrictions are attached to the deed rather than to the landowner (that is, they 'run with the land'), they avoid the transaction costs that would be necessary if each new resident had to negotiate anew with all existing residents. In this sense, land use servitudes represent an effective private alternative to zoning for small-scale developments. They are less effective, however, in controlling externalities in large-scale urban areas where development occurs in a piecemeal fashion over time.

Trespass and nuisance law also represent private alternatives to zoning. As noted above, trespass is effective in internalizing small-scale intrusions (for example, boundary disputes between neighbors), while nuisance law is best suited to harms that affect a few individuals (Ellickson, 1973). However, nuisance law is not well-suited to internalizing harms that are dispersed across a large number of landowners (public nuisances) because no one owner has an adequate incentive to incur the cost of bringing a nuisance suit, even though the aggregate harm may exceed the benefit (Landes and Posner, 1987, Chapter 2). Public regulation is the best remedy in these cases because the government can act as an agent for the group of affected landowners.

8. Public property and public use of private property

In section 2 we noted that public or state ownership was one of the primary types of property rights. Here we examine the rationale for public ownership and public control (including regulation and takings) of private property.

8.1. The optimal scale of ownership

When transaction costs are positive, the private ownership of land is not always the most efficient means of maximizing land value. The primary advantage of private ownership is that it creates the proper incentives for use and investment regarding actions taken within the boundaries of the property. However, since different uses of land have different optimal boundary requirements, it may be the case that the scale of an activity

[129] There is almost no economic analysis of servitudes or land use doctrines; but see Stake (1998b).

exceeds the existing boundaries of ownership (Ellickson, 1993). For example, Coase's example of straying cattle suggests that the rancher's parcel was too small. One solution to this problem is contracting between ranchers and neighboring owners who suffer harm (Ellickson, 1991), but if contracting costs are high, a better solution may be to consolidate ownership of the parcels (Libecap, 1989). In this way, market transactions are replaced by internal governance methods (Ostrom, 1990).

The optimal solution depends on the cost of contracting among landowners (which increases with greater decentralization) compared to the cost of governance under consolidated ownership.[130] Both types of costs are likely to increase with the scale (or size) of the asset, because reaching an agreement will generally require dealing (either through contracts, monitoring, or political deals) with larger and more heterogeneous group of parties. For the largest scale activities, state ownership will likely dominate both private ownership (because no single owner will want to hold and manage such a large, undiversified portfolio of assets), and private contracting (because the state can use decision rules that do not require unanimity to lower the costs of agreement).[131] In an empirical application Lueck (1989, 1995a) examines the ownership regimes that govern wildlife, a resource that often has an optimal scale of management that far exceeds the typical boundaries of private land holdings. Lueck finds a mix of private contracting and government ownership regimes that have developed in response to the potential externality problems.

8.2. *The public trust doctrine*

The public trust doctrine is an ancient doctrine which grants ownership of navigable rivers, shorelines, and the open sea to the public.[132] The public trust doctrine can be viewed as the judicial creation of common property, which has roots in Roman law and the English common law. English and Roman public trust law both acknowledged inalienable public rights in navigable waterways and the foreshore. They allowed, for example, unrestricted access to large watercourses for travel and transportation. The public trust doctrine also has been a part of American law, providing public access to navigable waterways and authorizing state control over tidelands.[133] In essence, the public trust doctrine 'defines an easement that members of the public hold in common' (Huffman, 1989, p. 527), thus creating a sort of common property resource among a disorganized public. In recent years some courts have extended the doctrine into new areas—mostly environmental assets—such as beaches, lakes, stream access, and

[130] The problem is analogous to Coase's (1937) theory of the optimal boundary between the market and the firm.

[131] Libecap (1989), however, notes that virtually the same forces that make private contracting costly also make political solutions costly.

[132] For an introduction see Dukeminier and Krier (2002, pp. 816–823).

[133] The seminal case is *Illinois Central Railroad v. Illinois*, 146 U.S. 387 (1892).

wildlife (Sax, 1970). For example, *National Audubon Society v. Superior Court*,[134] perhaps the most important modern case, extended public trust status to wildlife habitat at California's Mono Lake, thereby effectively reallocating water rights. Similarly, recent law in Montana has extended the public trust claim to recreational uses (e.g., fishing, rafting) of waterways.[135]

In its traditional application, navigable waters, the public trust asset was essentially a public good. When an asset is a public good, unrestricted access will not cause dissipation from overuse of the resource.[136] On the other hand, when the resource has private good characteristics, unrestricted access (especially by a large number of users) can trigger the rule of capture and create a classic open access problem. Indeed, some critics (Cohen, 1992; Huffman, 1989) of new environmental applications of the public trust doctrine argue that expanding access to resources will lead to their degradation through overuse.[137] For instance, a public trust conversion of a private beach into a public beach may well lead to crowding and pollution of the beach.

8.3. Eminent domain and regulatory takings

Large-scale economic developments like dams and irrigation projects, railroads, highways, and shopping centers often involve the assembly of a large contiguous parcel of land from relatively small and separately owned parcels. In all of these cases, the provider, whether public or private, faces a potential holdout problem (Cohen, 1991; Strange, 1995). The source of this problem is that, once assembly becomes public knowledge, each landowner realizes that he or she can impose a substantial cost on the provider by refusing to sell. This knowledge confers monopoly power on owners, who can each hold out for prices in excess of their true valuations, thereby endangering completion of the project.[138]

One solution to the land assembly problem is to allow forced sales—that is, replace property rule protection of each owner's land with liability rule protection.[139] This is the economic justification for the eminent domain power of the state (Posner, 2003, p. 55) which has common law origins. The 'Takings' clause of the Fifth Amendment of the

[134] 658 P.2d 709 (Cal. 1983).

[135] *Montana Coalition for Stream Access v. Curran*, 682 P.2d 161.

[136] There is still the problem of raising revenues to police and maintain the asset.

[137] Cohen (1992) and Rose (1986) note how an expansive public trust doctrine can be used by governments to avoid the Constitution's takings clause.

[138] In this sense, the holdout problem resembles the anti-commons problem discussed in section 2.2. It is important to distinguish this problem from the case of single owners of dispersed parcels who seek the best price for their property in one-on-one transactions. This is not a holdout problem because the owners are not seeking a price above the true valuation of their property, nor does any one owner's refusal to sell affect the transfer of other parcels.

[139] One well known example of forced transfers is state law that compels the formation of reservoir-wide conservation 'units' for oil and gas production. Similar laws govern irrigation districts, soil conservation and predator control districts.

U.S. Constitution explicitly grants the power, saying 'nor shall private property be taken for public use, without just compensation.' The key components of this clause are the requirements of 'public use' and 'just compensation,' which we discuss in turn.

8.3.1. Public use of private property

Merrill (1986) examines the scope of the takings power in the context of the public use requirement. He draws a distinction between the 'means' and 'ends' approach to public use. The means approach concerns the manner in which land is acquired for large-scale projects (is there a holdout problem?), while the ends approach refers to the use of the land (is it for a public or private good?). It is important to note that these are separable categories—that is, not all public goods require land assembly, and some private goods do. According to the ends approach, the takings power should be limited to provision of public goods by the government, whereas according to the means approach, it should be granted to any provider facing a holdout problem.[140]

Merrill's 'ends approach' appears more consistent with the plain meaning of public use, but it potentially results in two types of 'errors.' First, it may result in the use of eminent domain for the provision of public goods not requiring land assembly (Fischel, 1995a, p. 74). Merrill argues, however, that this overuse of the takings power (i.e., the substitution of coercive for consensual transactions) is self-limiting in the sense that the costs of market acquisition are generally less than the costs of eminent domain. Second, the ends approach apparently denies use of eminent domain to private providers facing a holdout problem. Historically, however, courts have tended to act in accordance with the means approach by granting takings power to private parties like railroad and canal builders who face serious holdout problems, though they nearly always attempt to justify their action in terms of the ends approach—that is, they identify some public benefit from the project (Merrill, 1986, p. 67). The need for such justification is somewhat surprising, however, given that courts routinely use liability rules (i.e., money damages) as a remedy in other disputes involving private parties. For example, awarding damages to the plaintiffs in the *Boomer* case rather than shutting the factory down amounted to a 'private taking' by the factory (Fischel, 1995a, p. 76). This was appropriate, we argued, because the factory faced a kind of holdout problem. The point is that the actual use of eminent domain appears to reflect economic logic (the means approach), and when necessary, courts bend the meaning of public use to conform to this standard (Fischel, 1995a, pp. 75–77). However, some critics have argued that in recent years, the public use doctrine has expanded to include private development projects seemingly beyond the original intention of the doctrine.[141]

[140] Ulen (1992) argues that eminent domain should only be used when both conditions are met.

[141] See, for example, *Kelo v. City of New London*, 125 S.Ct. 2655 (2005), in which the Supreme Court permitted the use of eminent domain to transfer land from one private party to another as part of a comprehensive redevelopment plan. The Court held that the general benefits to the community from economic growth satisfied the public use requirement. Also see *Poletown Neighborhood Council v. City of Detroit*, 410 Mich. 616, 304 N.W.2d 455 (1981).

8.3.2. Just compensation for takings

In addition to public use, the eminent domain clause requires payment of just compensation following a taking. Courts have interpreted this to mean 'fair market value.' Several authors have argued, however, that fair market value generally under compensates landowners because it ignores the owner's subjective value (e.g., Knetsch and Borcherding, 1979; Fischel, 1995b). Since subjective value is part of the opportunity cost of a taking, failure to compensate for it potentially results in over acquisition of land by the government. In an empirical study of land acquisition in Chicago, Munch (1976) found that compensation amounts differed systematically from market value. Generally, owners of high valued properties were overcompensated, while owners of low valued properties were under compensated. Epstein (1985a, Ch. 15) however, contends that taxes used to finance compensation are themselves a form of taking, which act as a limit on the amount of land taxpayers will permit the government to acquire. In the same vein, Fischel (1995a, p. 211) argues that market value may be the 'proper' measure of just compensation because it balances the cost of undercompensation against the higher taxes that full compensation would require.

The economic literature on takings has focused on the link between compensation and investment decisions of landowners. One of the primary contributions of this literature, initiated by Blume, Rubinfeld, and Shapiro (BRS) (1984), has been to show that it may be inefficient to pay *any* compensation. This 'no compensation result' can be illustrated by a simplified version of the BRS model. Suppose there are multiple parcels of land, each worth $V(x)$ if the landowner makes an irreversible investment x, where $V' > 0$ and $V'' < 0$. The land may also be valuable for public use, yielding a benefit of $B(y)$, where y is the number of parcels taken and $B' > 0$, $B'' < 0$ (Fischel and Shapiro, 1989). (Alternatively, y may be interpreted as the probability of a taking, or the fraction of a given parcel's value that is extinguished by a regulation—see section 8.3.3 below.) In the event of a taking, suppose that compensation of $C(x)$ will be paid for each parcel taken, where $C(x) \geq 0$, and $C' \geq 0$. Thus, total compensation will be $yC(x)$.

The time sequence is that landowners choose x given the anticipated behavior of the government and the compensation rule; then the government chooses y and pays $C(x)$. We will assume various objective functions for the government below. As a benchmark, note that the first-best choices (x^*, y^*) maximize $B(y) + (1 - y)V(x) - x$, which is the sum of private and public benefits. The relevant first-order conditions are

$$(1 - y)V'(x) - 1 = 0 \tag{8.1}$$

$$B'(y) - V(x) = 0. \tag{8.2}$$

Now consider the decisions separately made by each party. In the first scenario, we view the government's taking decision as exogenous—that is, it is unaffected by the compensation rule. This is the assumption BRS (1984) make in their basic model, and represents what Fischel and Shapiro (1989) refer to as an 'inexorable' government. In this case, y is fixed (so condition (8.2) is irrelevant), while the landowner chooses x to maximize $(1-y)V(x)+yC(x)-x$, which must satisfy

$$(1-y)V'(x)+yC'(x)-1=0. \tag{8.3}$$

Let x^l be the solution. Comparing (8.3) to (8.1) shows that $C'=0$ is necessary for the landowner to invest efficiently; that is compensation must be lump sum to ensure that $x^l=x^*$ (BRS, 1984). Intuitively, any positive relationship between x and the amount of compensation creates moral hazard that results in over-investment. It immediately follows that no compensation ($C(x)\equiv 0$ for all x) is efficient, although any lump sum rule is consistent with efficiency.[142]

The efficiency of zero compensation does not hold up, however, under different assumptions about the government's behavior. Suppose, for example, that the government chooses y to maximize social welfare. Such a government has been characterized as 'benevolent' (Hermalin, 1995) or 'Pigovian' (Fischel and Shapiro, 1989). The optimal choice of y in this case is given by the first-order condition in (8.2). Note that, because the government chooses y after the landowner's investment of x is in place, (8.2) defines a function $y^g(x)$, where[143]

$$\frac{\partial y^g}{\partial x}=\frac{V'}{B''}<0. \tag{8.4}$$

The amount of land taken is decreasing in x because the more the landowner has invested, the higher is the opportunity cost of a taking.

The landowner's objective function is the same as above, but he now maximizes it subject to the anticipated behavior of the government as described in (8.4). The first-order condition is

$$(1-y)V'(x)+yC'(x)-[V(x)-C(x)](\partial y^g/\partial x)-1=0. \tag{8.5}$$

Note that compensation must again be lump sum, but zero compensation is no longer consistent with efficiency. This is reflected by the third term in (8.5), which implies that the landowner will over-invest if $C(x)<V(x)$ and under-invest if $C(x)>V(x)$. Intuitively, if the landowner expects to be under compensated in the event of a taking, he will increase his investment in order to lower the probability of a taking. Conversely, if he expects to be overcompensated, he will under-invest in order to raise the probability of a taking (Miceli, 1991; Hermalin, 1995).

This version of the model embodies two potential sources of moral hazard for the landowner. The first is the tendency to over-invest if compensation is an increasing function of x (the basis for the BRS no-compensation result), and the second is the effect of x on the government's taking decision through (8.4). One compensation rule that resolves both problems and induces a first-best level of investment is $C=V(x^*)$.[144]

[142] This is another example of the *paradox of compensation* (see section 7.3).

[143] By definition, $y^g(x^*)=y^*$.

[144] This rule is not the only one that achieves the efficient outcome. One alternative will be discussed in section 8.3.3 below, and Hermalin (1995) proposes others.

That is, compensation is set at the full value of the land, evaluated at the efficient level of investment.

It is important to emphasize that the justification for compensation in this version of the model is *not* to prevent excessive acquisition of land by the government, as is often argued. Rather, it arises from the sequential decisions of the parties. Suppose, however, that the government is not benevolent, but instead acts on behalf of the majority, or some group with political influence (those who receive the benefits of the taking) while ignoring the costs to individual property owners, except to the extent that it must pay them compensation (Fischel and Shapiro, 1989; Hermalin, 1995; Nosal, 2001). Such a government is said to have 'fiscal illusion' in that only dollar costs enter its cost-benefit calculation (BRS, 1984).

In this case, the government chooses y to maximize $B(y) - yC(x)$, which yields the first-order condition

$$B'(y) - C(x) = 0. \tag{8.6}$$

As before, this defines a function $\hat{y}(x)$ whose characteristics depend on the nature of the compensation rule. The landowner now maximizes his objective function subject to $\hat{y}(x)$, which yields the first-order condition in (8.5) with $\partial y^g/\partial x$ replaced by $\partial \hat{y}/\partial x$. Clearly, $C = V(x^*)$ will again induce first-best investment by the landowner based on the same reasoning above. Moreover, setting $C = V(x^*)$ in (8.6) also yields the efficient taking decision by the government.[145]

A final argument against the no-compensation result is due to Michelman (1967), who argues that compensation should depend on a comparison of the 'settlement costs' of paying compensation with the 'demoralization costs' of not paying compensation. In terms of the preceding analysis, settlement costs include administrative costs and the costs associated with moral hazard, while demoralization costs arise from the risk of an uncompensated taking (Fischel and Shapiro, 1988; Fischel, 1995a, Chapter 4). Compensation should therefore be paid when the demoralization costs exceed the settlement costs. A related justification for compensation is that it provides risk averse landowners public insurance against the political risk (demoralization costs) associated with takings, given that private insurance for such risk is not readily available (Blume and Rubinfeld, 1984; Kaplow, 1986; Rose-Ackerman, 1992.[146]

[145] Alternatively, Fischel and Shapiro (1989) consider a compensation rule of the form $C = sV(x)$ where s is the fraction of the value of the land that the government will pay in the event of a taking. They argue that this is an easier rule to administer compared to $C = V(x^*)$ because it does not require the government to calculate x^*. The shortcoming is that the optimal value of s, which is strictly between zero and one, only achieves a second-best outcome.

[146] Though it has not been the subject of economic models, the public use constraint is empirically important and might be examined formally through the $B(y)$ function. For example, if $B(y) = B(\sum y_i)$, where $i = 1, \ldots n$ (and n is the entire population) then a public use requirement could limit state taking to those cases where some supermajority fraction—say $(n - p)/n$, where $p < n$—was required (implicitly in the doctrine).

8.3.3. Regulatory takings

To this point we have focused on compensation for takings of land under eminent domain, or physical acquisitions, but much more common are government regulations that restrict land uses without actually depriving the owner of title. Examples include zoning laws, and environmental and safety regulations. Historically, courts have granted the government broad powers to enact such regulations as an exercise of its police power, but when a regulation becomes especially burdensome, the affected landowner may claim that a 'regulatory taking' has occurred and seek compensation under the eminent domain clause.[147]

The case law on this question is extensive, and though the Supreme Court has advanced several tests for compensation, there is no consensus on when a regulation crosses the threshold separating a non-compensable police power action from a compensable taking. Some noteworthy tests are the *noxious use doctrine*, which says that a regulation is not compensable if it protects the 'health, morals, or safety of the community;'[148] the *diminution of value test*, which says that compensation is due if a regulation 'goes too far' in reducing the value of the regulated land;[149] and the *nuisance exception*, which says that compensation is not due for regulations that prevent activities that would be classified as nuisances under the governing state's common law.[150]

Like the takings analysis above, the trade-off for regulatory takings concerns the efficiency of the land use decision on the one hand, and the regulatory decision on the other. As a way of examining the threshold nature of this choice, consider the following compensation rule (Miceli and Segerson, 1994, 1996):

$$C = \begin{cases} 0, & \text{if } y \leq y^* \\ V(x), & \text{if } y > y^*. \end{cases} \tag{8.7}$$

Here, we can interpret y to be the extent of a landowner's value that is lost as a result of a regulation. Note that this rule is conditional on the behavior of the government in that it requires full compensation if the government over-regulates ($y > y^*$), but requires no compensation otherwise ($y \leq y^*$). In this sense, it is like a negligence rule in tort law, and it yields and efficient equilibrium for the same reason.[151]

[147] Such claims take the form of so-called inverse condemnation suits.

[148] *Mugler v. Kansas*, 123 U.S. 623, 1887.

[149] *Pennsylvania Coal v. Mahon*, 260 U.S. 393, 1922.

[150] *Lucas v. South Carolina Coastal Council*, 112 S.Ct 2886, 1992.

[151] The proof of efficiency is complicated, however, by the fact that the landowner and regulator act in sequence. Note first that if $x = x^*$, the government's optimal response is y^*. To see why, observe that it will never choose $y < y^*$ given $B' > 0$, and it will prefer y^* to $y > y^*$ since $B(y^*) > B(y^*) - y^*V(x^*) \geq \max_{y>y^*} B(y) - yV(x^*)$. Next, suppose $x > x^*$. The government's optimal response in this case is also y^* since $y^g(x) < y^g(x^*) \equiv y^*$ for $x > x^*$ by (8.4), which again implies that $B(y^*) > B(y^g(x)) - y^g(x)V(x) \geq \max_{y>y^*} B(y) - yV(x)$. Finally, if $x < x^*$, the government will choose y^* if x is near x^*, but it will prefer $y > y^*$ if x is sufficiently small. If the government is expected to choose y^*, $C = 0$ and the landowner

In addition to the efficiency of (8.7), the rule advances our understanding of actual takings law in several ways. First, it allows us to interpret 'noxious uses' as those activities that are efficiently regulated, and for which no compensation is due (the top line of (8.7)). Second, it provides an economic standard for when a regulation 'goes too far' under the diminution of value test. Specifically, a regulation goes too far—and compensation is due—when it is inefficiently enacted (the second line of (8.7)).[152] Finally, it establishes a standard that is economically equivalent to the common law definition of a nuisance (an activity that is efficiently prohibited),[153] and hence is consistent with the threshold for compensation implied by the nuisance exception.

8.3.4. Compensation and the timing of development

It is clear from the above discussion that regulations often redefine property rights to the disadvantage of landowners. Faced with the threat of no compensation for alterations of their property rights, landowners can often reclaim these rights because they have private information and a first mover advantage over regulatory agencies and legislatures. In the process, they can preempt regulations and may do so in ways that counter the intended goals of the regulations. Land preservation and environmental regulations are perhaps the classic cases (Cohen, 1999; Dana, 1995). While regulators consider restrictions to preserve land, developers race to beat the regulations, often leading to more rapid development than would have otherwise occurred.

The incentive to preemptively develop can be seen in a two-period model of a landowner and a regulatory agency which can invoke a land use regulation that will lower the value of the land by preserving some environmental amenity (e.g., endangered species habitat, open space).[154] The land's value under the regulation depends on the landowner's behavior. Specifically, the landowner can choose to maintain (m) or destroy (d) the amenity in period 1. The landowner has private information about the amenity and has a clear first mover advantage over the agency because of this information and because of his ownership incentives. Development and thus destroying the

maximizes $(1 - y^*)V(x) - x$, which yields x^*. This leaves $x < x^*$ and $y > y^*$ as the only possible outcome besides (x^*, y^*), but for this to be an equilibrium, it must be true that $B(y) - yV(x) > B(y^*)$ for the government, and $V(x) - x > (1 - y^*)V(x^*) - x^*$ for the landowner. Summing these conditions implies $B(y) + (1 - y)V(x) - x > B(y^*) + (1 - y^*)V(x^*) - x^*$, which contradicts the definition of x^* and y^*. This proves that $x < x^*$ and $y > y^*$ cannot be an equilibrium.

[152] In this sense, the noxious use doctrine and the diminution of value test are two sides of the same coin. This interpretation suggests that the disagreement between Holmes and Brandeis in *Pennsylvania Coal* was over *facts* rather than *law* (Miceli and Segerson, 1996, p. 70).

[153] Keeton et al. (1984, p. 630) defines a nuisance as an activity for which 'the amount of the harm done outweighs the benefits.'

[154] The model here follows Lueck and Michael's (2003) application to the federal Endangered Species Act (ESA). Miceli and Segerson (1996) present a similar model of development with irreversible investment that generates premature development without compensation. Innes, Polasky, and Tschirhart (1998) also examine the incentives for landowners under the ESA.

amenity has a one-time cost (K_D) and generates benefits (B_D) from development. K_D is the cost of developing early, for example, harvesting timber before it has reached the optimal harvest age. If the amenity is destroyed the probability that the land will be regulated is zero. However, if the amenity is maintained there is a probability, $\gamma \in (0, 1)$, that the regulation will be invoked because the agency will deem the amenity worth preserving.

If the regulation is invoked, the landowner loses all benefits from development in period 2, but he may earn a smaller amount of benefits from an alternative land use that does not harm the amenity ($B_A < B_D$). In the absence of the regulation, the optimal time to develop is in period 2. This is true both because it is privately optimal for the owner to wait to avoid K_D, and because the amenity may be preserved. The landowner takes as given market prices (which determine the magnitudes of the various benefits and costs) and the probability the agency will invoke the regulation.

The landowner will maximize the expected value of the land by choosing to destroy the amenity if the expected value of early development exceeds that of waiting, or if $(B_D - K_D) > (1 - \gamma)B_D + \gamma(B_A + C)$, where $C \geq 0$ is the expected compensation in the event of a regulation. This inequality leads to several straightforward comparative static predictions. First, if $C = 0$, increases in the probability that the land will be regulated (γ) will increase the probability of preemptive development. Second, as the net value of development ($B_D - B_A$) increases, amenity destruction is more likely. Third, as the opportunity cost of early development increases (K_D) it is less likely that habitat destruction will occur. Finally, preemptive development becomes less likely if $C > 0$, and full compensation ($C = B_D - B_A$) deters early development with certainty.

Dana (1995) offers anecdotal evidence of such preemptive development in the absence of compensation, and Lueck and Michael (2003) find that the federal Endangered Species Act (ESA) has led some forest landowners to preemptively harvest timber in order to avoid costly land-use restrictions. For example, they find that landowners in North Carolina who are closer to populations of endangered red-cockaded woodpeckers (and thus subject to potentially costly timber harvest restrictions) are more likely to prematurely harvest their forest and choose shorter forest rotations. In this setting the empirical evidence indicates some endangered species habitat has been reduced on private land because of the ESA's land use regulations. The extent of such counter-productive regulations is not widely known and is a potentially important area of empirical research.

9. Inalienability of property rights

As Posner (2003, p. 75) notes: 'the law should, in principle, make property rights freely transferable' in order to allow resources to move to their most highly valued uses and to foster the optimal configuration of assets. He further notes that the long term trend has been to do just this: '[The] history of English land law is a history of efforts to make land more easily transferable and hence to make the market in land more efficient.' The same is true for most assets, including human capital. Yet, there remain many restrictions in both the common law and in statutes and regulations that limit the alienability of property. These 'inalienability' rules (Calabresi and Melamed, 1972) typically take the form of government restrictions on property and hence should be distinguished from state property in that they do not suspend or replace private ownership of the property in question, but merely limit its alienability (only one of the bundle of property rights). Inalienability rules apply to body parts, children, voting, military service, cultural artifacts, endangered animal species, the right to freedom (laws against slavery), certain natural resources, state property (as noted in section 2.4) and many other cases.[155] These restrictions may be complete (e.g., slavery) or and partial (e.g., water transfers). And, of course, they are enforced in varying degrees, both as part of law and as part of group-based rules such as those that arise to govern common property. There is little empirical literature that tests various theories of inalienability so the discussion here is mostly limited to claims of plausibility and consistency with previous analysis in this chapter.

The dominant economic reason for restrictions on alienability is that externalities can arise from transfers (Barzel, 1997; Epstein, 1985b; Rose-Ackerman 1992, 1998; Posner, 2003).[156] Transfers can have external effects on third parties if the rights to the assets are not well-defined.[157] As noted in the discussion of first possession rules, well-defined rights mean that exclusive rights are defined to the stock and, accordingly, its stream of flows over time. This implies that the law should allow rights transfers when there is clear ownership of resource stocks. When rights to the stock remain ill-defined, however, there may be a rationale for limiting, even prohibiting, transfers of the claimed flows in order to protect the asset itself. For example, the widespread prohibition on trade in wild game is likely to be such a case (Lueck, 1989, 1998), though even here limits on markets can potentially deter the formation of property rights as discussed in section 4. Restrictions on the sale of children may find a rationale for the same reason: a market for children (or game) would lead to 'poaching' of animals and kids for which property rights enforcement is extremely costly.[158]

The law on western water transfers offers a useful example of how less extreme imperfections in property rights can lead to restrictions on transfers that actually serve to clarify rights (Barzel, 1997; Lueck, 1995a, 1995b). The doctrine of prior appropriation, found in most western states, allows ownership of water separate from land ownership. Owners of such water rights (the name stems from the first possession rule used to

[155] Rose-Ackerman (1998) discusses political rights in detail but we do not examine these here. Andolfatto (2002), in work disconnected from the economics of law, shows how limits on transfers can be efficiency enhancing in the context of social programs such as social security entitlements.

[156] As usual the legal literature has discussed 'distributive justice' reasons for restrictions on transfers but we do not examine those here. Rose-Ackerman (1985) links these arguments to economic arguments.

[157] We explicitly ignore market effects from trade or 'pecuniary externalities.'

[158] Although we have argued that liability rules can often internalize externalities when transaction costs are high, the examples suggest that some potential harms may best be dealt with by banning certain activities altogether.

originally assign them) are generally free to use and transfer these rights, but certain restrictions on transfers are common, most notably prohibitions on transfers to parties who are not located in the stream basin or who will have different uses for the water. On the surface these restrictions seem blatantly inefficient, akin to restricting the sale of milk to someone who lives in the same city or uses it only on their cereal like everyone else. Yet a consideration of the way rights to water are defined illuminates these restrictions. Assume the water is diverted for irrigation, and that a portion of the water is returned to the stream after irrigation (the 'return flow'). Only the water actually used or 'consumed' is valued by the users, so ideally water rights would be defined in terms of actual water consumed. Under a perfect system of ownership, the transfer of water rights would maximize the value of the water in the stream by generating an equilibrium in which the marginal value of water is the same for all users (Johnson, Gisser, and Werner, 1981).[159]

In practice, however, it is too costly to measure and define water according to consumptive use, so typically only the amount diverted is actually measured and traded.[160] This implies that a consensual trade for diverted water rights might adversely impact a downstream water user if the trade alters the amount of water actually consumed. The return flow will vary depending on the nature and location of the water's diversion and use. If the water is diverted out of the basin then there is absolutely no return flow and a transfer out of basin will impose externalities on other rights holders not party to the transfer. If water is used more intensively by a new user, the return flow will be lower though not zero. Thus, if water is defined over diverted use, then unconstrained water rights transfer will not maximize the value of the water. But if transfers are restricted so that diverted rights mimic consumptive rights, then this restricted rights regimes will indeed maximize the value of the water. Restrictions on use and on out-of-basin transfers seemingly meet these conditions and can thus be explained as efficient restrictions on rights.[161] More generally, when an asset is complex—water rights can be defined over diversion, consumption, and even quality for that matter—then efficient property rights regimes may contain restrictions on alienation that might seem to be inefficient. Only by moving away from a strict neoclassical view of property rights can these restrictions be seen as having an economic rationale.

[159] Let n be the number of rights holders, C_i the consumptive use of for user i, W the stock of water in the basin, and f_i the marginal value of consumed water. The social problem is to maximize $V = \sum_{i=1}^{n} \int_{0}^{c_i} f_i'(w_i)dw_i$ subject to the water stock constraint $\sum_{i=1}^{n} C_i = W$. This solution requires $f_1' = f_2' = \cdots = f_n' = \lambda$ where λ is the Lagrange multiplier and equal to the marginal value of consumed water.

[160] New Mexico, however, does defines rights in terms of consumption and has generally fewer transfer restrictions (Johnson, Gisser, and Werner, 1981).

[161] Now let diversion by user i be D_i where $C_i \equiv D_i(1 - F_i)$ where F_i is user i's fraction of the water returned to the stream. The stock constraint now becomes $\sum_{i=1}^{n}[(1 - F_i)D_i] = W$ and the new solution requires $f_1'/(1 - F_1) = f_2'/(1 - F_2) = \cdots = f_n'/(1 - F_n) = \lambda$. If transfers are restricted so that $F_i = F_j$ for all $i \neq j$ users, then this condition is identical to the first-best value derived above. Note that neither of these water models examines the costs of measurement and enforcement.

Another reason for restriction on transfers is asymmetric information, particularly that leading to adverse selection (Rose-Ackerman, 1998). In the worst case, adverse selection can completely dry-up all markets associated with a commodity where product quality cannot be observed prior to purchase (Akerlof, 1970). Some have used this argument to explain bans on the sale of human blood and body parts, while allowing donations (a modified form of inalienability). It is not clear, however, how donations rather than sales will eliminate the adverse selection problem. Friedman (2000, p. 242) argues that a better reason for only allowing organ donations is to discourage kidnapping and murder for purposes of harvesting and selling organs. Friedman's point is similar to that made above regarding poaching. Laws against involuntary slavery would similarly discourage forced capture and sale of people into slavery (Barzel, 1977). (It is harder to justify laws against 'voluntary slavery,' or indentured servitude, as a penalty for loan default when borrowers lack financial assets to serve as collateral.)

Similar restrictions on the types of property servitudes allowed (e.g., limits on 'negative and in gross' easements) might be explained based on asymmetric information (Dnes and Lueck, 2007). In legal studies, the limitations on servitudes have been explained in relation to the Rule against Perpetuities as a method of preventing 'clogging title' (e.g., Gray and Gray, 2000), although this argument has not been explored in economics. Consider the market for land of two types: fee simple unencumbered and land encumbered with a generic servitude. Assume that only the seller of the plot knows whether or not the land is encumbered. Buyers do not have this information but only know that one-half of the land is encumbered. The value of an unencumbered plot is V^f, while the value of the encumbered plots is $V^s < V^f$. Given the information asymmetry buyers will only pay the expected value of a plot, $EV = (V^s + V^f)/2 < V^f$. Following Akerlof (1970) and related literature, this means there will be no market equilibrium for the unencumbered plots; that is, only 'low quality' encumbered plots will be present in the market. Institutions that provide information (e.g., recording and registration systems) could eliminate asymmetry, as could some institutional 'rules of the game'.[162] Posner (2003, p. 75) offers a similar reason for the use of terminable easements for railroads rather than fee simple ownership to a narrow strip of land passing by thousands of other owners. Parisi, Schulz, and Depoorter (2006), however, argue that anti-commons incentives can also explain such common limitations on property regimes.

Two other factors may be important in determining restrictions on transferability. First, interest group pressure may lead to restrictions on transfers that have purely redistributive effects. The literature on economic regulation provides evidence on this from many commercial areas. One area where such interest group pressure has been important has been the broadcast spectrum (Hazlett 1990, 1998). There the restriction on use and transfers seem to have little rationality in limiting externalities or mitigating information problems, but instead serve to protect incumbent users from competition. Second, for state property administered by bureaucratic agencies, limits on transfers

[162] This adds an additional rationale for title and recording systems discussed earlier.

may serve to limit the potential moral hazard of the bureaucrats who might gain from transfers without facing the opportunity cost of the transfer.

10. Conclusion

The economic analysis of property rights and the economic analysis of law are the twin offspring of Coase's (1960) seminal work. Yet, today the economics of property law is a poor cousin to the economics of contracts, torts, and many other areas. In part this is because economic analysis of property law has not been as welcome among property law scholars as it has been among legal scholars of antitrust, contracts and torts. In part is it because property law is so broad, making comprehensive analysis a daunting task.

In this chapter we have surveyed the somewhat disjoint literature developed by economists and legal scholars, elaborated on some of the basic models, and highlighted areas where more work remains to be done. While many important issues remain it can be claimed that economic analysis reveals a fundamental logic to the main doctrines and features of property law. In short, the observed structure of property rights and property law can be best understood as a system of societal rules designed to maximize social wealth. Among the most important remaining issues for study is a systematic analysis of how the law addresses the use and transfer of complex assets. And, as always, more detailed empirical work is needed to fully test and understand the rationale for the law and its effects.

Acknowledgements

We are grateful for the comments of Doug Allen, Tom Hazlett, Peter Menell, Eric Rasmusen, Steve Shavell, and Ian Wills, as well as the participants in the Law and Economics Handbook Conference at Stanford Law School, the European School of New Institutional Economics, and a workshop at the University of Michigan. Lueck was supported by the Cardon Endowment for Agricultural and Resource Economics; Miceli was supported by a Grillo Family research grant.

References

Acheson, J.M. (1988). The Lobster Gangs of Maine. University Press of New England, Hanover, NH.

Ackerberg, D., Botticini, M. (2002). "Endogenous matching and the empirical determinants of contract form". Journal of Political Economy 110, 564–591.

Agnello, R.J., Donnelley, L.P. (1975). "Property rights and efficiency in the oyster industry". Journal of Law and Economics 18, 521–534.

Akerlof, G. (1970). "The market for 'lemons': Quality Uncertainty and the Market Mechanism". Quarterly Journal of Economics 84, 488–500.

Alchian, A.A. (1965). "Some economics of property rights". Il Politico 30, 816–829. Reprinted in Alchian, A.A. (Ed.) (1977). Economic Forces at Work. Liberty Press, Indianapolis.

Allen, D.W. (1991). "Homesteading and property rights; or, 'How the west was really won'". Journal of Law and Economics 34, 1–23.

Allen, D.W. (1998). "Cropshare contracts". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 1. Stockton Press, New York, pp. 569–572.

Allen, D.W. (1999). "Transaction costs". In: Bouckaert, B., DeGeest, G. (Eds.), Encyclopedia of Law and Economics, vol. I. Edward Elgar, Cheltenham, pp. 893–926.

Allen, D.W. (2002). "The rhino's horn: incomplete property rights and the optimal value of an asset". Journal of Legal Studies 31, S339–S358.

Allen, D.W., Lueck, D. (1995). "Risk preferences and the economics of contracts". American Economic Review 85, 447–451.

Allen, D.W., Lueck, D. (1999). "The role of risk in contract choice". Journal of Law, Economics, and Organization 15, 704–736.

Allen, D.W., Lueck, D. (2003). The Nature of the Farm: Contracts, Risk and Organization in Agriculture. MIT Press, Cambridge, MA.

Alston, L., Libecap, G., Schneider, R. (1996). "The determinants and impact of property rights: land title on the Brazilian frontier". Journal of Law, Economics & Organization 12, 25–61.

Alston, L.J., Schapiro, M. (1984). "Inheritance laws across colonies: causes and consequences". Journal of Economic History 44, 277–287.

Anderson, T.L., Hill, P.J. (1975). "The evolution of property rights: a study of the American West". Journal of Law and Economics 18, 163–179.

Anderson, T.L., Hill, P.J. (1990). "The race for property rights". Journal of Law and Economics 33, 177–197.

Anderson, T.L., Lueck, D. (1992). "Land tenure and agricultural productivity on Indian reservations". Journal of Law and Economics 35, 427–454.

Andolfatto, D. (2002). "A theory of inalienable property rights". Journal of Political Economy 110, 382–393.

Arruñada, B. (2003). "Property enforcement as organized consent". Journal of Law, Economics and Organization 19, 401–444.

Bailey, M.J. (1992). "Approximate optimality of aboriginal property rights". Journal of Law and Economics 35, 183–198.

Baird, D., Jackson, T. (1984). "Information, uncertainty, and the transfer of property". Journal of Legal Studies 13, 299–320.

Baker, M., Miceli, T. (2005). "Land inheritance rules: theory and cross-cultural analysis". Journal of Economic Behavior and Organization 56, 77–102.

Baker, M., Miceli, T., Sirmans, C.F., Turnbull, G. (2001). "Property rights by squatting: land ownership risk and adverse possession statutes". Land Economics 77, 360–370.

Baker, M., Miceli, T., Sirmans, C.F., Turnbull, G. (2002). "Optimal title search". Journal of Legal Studies 31, 139–158.

Barzel, Y. (1968). "The optimal timing of innovations". Review of Economics and Statistics 50, 348–355.

Barzel, Y. (1977). "An economic analysis of slavery". Journal of Law and Economics 20, 87–110.

Barzel, Y. (1982). "Measurement cost and the organization of markets". Journal of Law and Economics 25, 27–48.

Barzel, Y. (1994). "The capture of wealth by monopolists and the protection of property rights". International Review of Law and Economics 14, 393–409.

Barzel, Y. (1997). Economic Analysis of Property Rights, 2nd edn. Cambridge University Press, Cambridge.

Barzel, Y., Kochin, L.A. (1992). "Ronald Coase on the nature of social cost as a key to the problem of the firm". Scandinavian Journal of Economics 94, 19–31.

Berger, L. (1985). "An analysis of the doctrine that 'First in time is first in right'". Nebraska Law Review 64, 349–388.

Besley, T. (1995). "Property rights and investment incentives: theory and evidence from Ghana". Journal of Political Economy 103, 903–937.

Besley, T. (1998). "Investment incentives and property rights". In: Newman, P. (Ed.), The New Palgrave Dictionary of Law and Economics, vol. 2. Stockton Press, New York, pp. 359–365.

Blackstone, W. (1766) [1979]. Commentaries on the Laws of England, of Book II. University of Chicago Press, Chicago.

Blume, L., Rubinfeld, D. (1984). “Compensation for takings: an economic analysis”. California Law Review 72, 569–624.

Blume, L., Rubinfeld, D., Shapiro, P. (1984). “The taking of land: when should compensation be paid?”. Quarterly Journal of Economics 99, 71–92.

Bohn, H., Deacon, R.T. (2000). “Ownership risk, investment, and the use of natural resources”. American Economic Review 90, 526–549.

Bostick, C.D. (1987). “Land title registration: an English solution to an American problem”. Indiana Law Journal 63, 55–111.

Bottomley, A. (1963). “The effect of common ownership of land upon resource allocation in Tripolitania”. Land Economics 39, 91–95.

Brooks, R., Murray, M., Salant, S., Weise, J.C. (1999). “When is the standard analysis of common property extraction under free access correct?”. Journal of Political Economy 107, 843–858.

Buchanan, J. (1983). “Rent seeking, noncompensated transfers, and laws of succession”. Journal of Law and Economics 26, 71–85.

Buchanan, J., Yoon, Y.J. (2000). “Symmetric tragedies: commons and anticommons”. Journal of Law and Economics 43, 1–14.

Bureau of Land Management, U.S. Department of the Interior (1999). Public Land Statistics 1999. U.S. Government Printing Office, Washington, D.C.

Calabresi, G., Melamed, A.D. (1972). “Property rules, liability rules, and inalienability: one view of the cathedral”. Harvard Law Review 85, 1089–1128.

Cannaday, R. (1994). “Condominium covenants: cats, yes; dogs, no”. Journal of Urban Economics 35, 71–82.

Cheung, S.N.S. (1969). The Theory of Share Tenancy. University of Chicago Press, Chicago.

Cheung, S.N.S. (1970). “The structure of a contract and the theory of a nonexclusive resource”. Journal of Law and Economics 13, 49–70.

Coase, R.H. (1937). “The nature of the firm”. Economica 4, 386–405.

Coase, R.H. (1959). “The federal communications commission”. Journal of Law and Economics 2, 1–40.

Coase, R.H. (1960). “The problem of social cost”. Journal of Law and Economics 3, 1–44.

Coase, R.H. (1988). The Firm, The Market and the Law. University of Chicago Press, Chicago.

Cohen, L.R. (1991). “Holdouts and free riders”. Journal of Legal Studies 20, 351–362.

Cohen, L.R. (1992). “The public trust doctrine: an economic perspective”. California Western Law Review 29, 239–276.

Cohen, M. (1999). “Monitoring and enforcement of environmental policy”. In: Tietenberg, T., Folmer, H. (Eds.), International Yearbook of Environmental and Resource Economics, vol. III. Edward Elgar, Cheltenham, pp. 44–106.

Coleman, J. (1988). Markets, Morals, and the Law. Cambridge University Press, Cambridge.

Cooter, R. (1985). “Unity in tort, contract, and property: the model of precaution”. California Law Review 73, 1–51.

Cooter, R., Ulen, T. (1999). Law and Economics, 3rd edn. Scott, Foresman and Co., Glenview, IL.

Cosgel, M., Miceli, T., Murray, J. (1997). “Organization and distributional equality in a network of communes: the shakers”. American Journal of Economics and Sociology 56, 129–144.

Dahlman, C.J. (1980). The Open Field System and Beyond: A Property Rights Analysis of an Economic Institution. Cambridge University Press, Cambridge.

Dana, D.A. (1995). “Natural preservation and the race to develop”. University of Pennsylvania Law Review 143, 655–708.

De Geest, G. (1992). “The provision of public goods in apartment buildings”. International Review of Law and Economics 12, 299–315.

Demsetz, H. (1967). “Toward a theory of property rights”. American Economic Review 57, 347–359.

Demsetz, H. (1972). “When does the rule of liability matter?”. Journal of Legal Studies 1, 13–28.

Demsetz, H. (1998). “Property rights”. In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 2. Stockton Press, New York, pp. 144–155.

De Soto, H. (2000). The Mystery of Capital. Basic Books, New York.

Dharmapala, D., Pitchford, R. (2002). "An economic analysis of 'riding to hounds': Pierson v. Post revisited". Journal of Law, Economics, and Organization 18, 39–66.

Dickinson, K. (1985). "Mistaken improvers of real estate". North Carolina Law Review 64, 37–75.

Dnes, A., Lueck, D. (2007). "Asymmetric information and the structure of servitude law", manuscript.

Dukeminier, J., Krier, J.E. (2002). Property, 5th edn. Aspen Law & Business, New York.

Dwyer, J., Menell, P. (1998). Property Law and Policy: A Comparative Institutional Perspective. The Foundation Press, Westbury, NY.

Eggertsson, T. (1990). Economic Behavior and Institutions. Cambridge University Press, Cambridge.

Eggertsson, T. (1992). "Analyzing institutional successes and failures: a millennium of common mountain pastures in Iceland". International Review of Law and Economics 12, 423–437.

Ellickson, R.C. (1973). "Alternatives to zoning: covenants, nuisance rules, and fines as land use controls". University of Chicago Law Review 40, 681–782.

Ellickson, R.C. (1986). "Adverse possession and perpetuities law: two dents in the libertarian model of property rights". Washington University Law Quarterly 64, 723–737.

Ellickson, R.C. (1989). "A hypothesis of wealth maximizing norms: evidence from the whaling industry". Journal of Law, Economics, and Organization 5, 681–782.

Ellickson, R.C. (1991). Order without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge.

Ellickson, R.C. (1993). "Property in land". Yale Law Journal 102, 1315–1400.

Ellickson, R.C., Been, V. (2000). Cases and Materials on Land Use Controls, 2nd edn. Aspen Law and Business, New York.

Epstein, R. (1979). "Possession as the root of title". Georgia Law Review 85, 970–990.

Epstein, R. (1985a). Takings: Private Property and the Power of Eminent Domain. Harvard University Press, Cambridge, MA.

Epstein, R. (1985b). "Why restrain alienation?". Columbia Law Review 85, 970–990.

Epstein, R. (1986). "Past and future: the temporal dimension in the law of property". Washington University Law Quarterly 64, 667–722.

Eswaran, M., Kotwol, A. (1985). "A theory of contractual structure". American Economic Review 75, 162–177.

Field, B. (1989). "The evolution of property rights". Kyklos 42, 319–345.

Fischel, W. (1985). The Economics of Zoning Laws: A Property Rights Approach to American Land Use Controls. Johns Hopkins University Press, Baltimore.

Fischel, W. (1995a). Regulatory Takings: Law, Economics, and Politics. Harvard University Press, Cambridge, MA.

Fischel, W. (1995b). "The offer/ask disparity and just compensation for takings: a constitutional choice perspective". International Review of Law and Economics 15, 187–203.

Fischel, W., Shapiro, P. (1988). "Takings, insurance, and Michelman: comments on economic interpretations of 'just compensation' law". Journal of Legal Studies 17, 269–293.

Fischel, W., Shapiro, P. (1989). "A constitutional choice model of compensation for takings". International Review of Law and Economics 9, 115–128.

Frech, H. (1979). "The extended Coase Theorem and long run equilibrium". Economic Inquiry 17, 254–268.

Friedman, D. (2000). Law's Order: What Economics has to do with the Law and Why it Matters. Princeton University Press, Princeton.

Fuedenberg, D., Gilbert, R., Stiglitz, J., Tirole, J. (1983). "Preemption, leapfrogging, and competition in patent races". European Economic Review 22, 3–31.

Gambetta, D. (1993). The Sicilian Mafia: The Business of Private Protection, 2nd edn. Harvard University Press, Cambridge, MA.

Geddes, R., Lueck, D. (2002). "The gains from self ownership and the expansion of women's rights". American Economic Review 102, 1079–1092.

Goetz, C., Scott, R. (1983). "The mitigation principle: toward a general theory of contractual obligation". Virginia Law Review 69, 967–1025.

Gordon, H.S. (1954). "The economic theory of a common property resource: the fishery". Journal of Political Economy 62, 124–142.

Gray, K.J., Gray, S. (2000). Elements of Land Law, 3rd edn. Butterworths, London.

Greif, A. (2006). Institutions and the Path to the Modern Economy. Cambridge University Press, Cambridge.

Haddock, D.D. (1986). "First possession versus optimal timing: limiting the dissipation of economic value". Washington University Law Quarterly 64, 775–792.

Hallwood, P., Miceli, T.J. (2006). "Murky waters: The law and economics of salvaging historic shipwrecks". Journal of Legal Studies 35, 285–302.

Hansen, Z., Libecap, G.D. (2004). "Small farms, externalities, and the dust bowl of the 1930s". Journal of Political Economy 112, 665–694.

Hansmann, H. (1991). "Condominium and cooperative housing: transactional efficiency, tax subsidies, and tenure choice". Journal of Legal Studies 20, 25–71.

Hansmann, H., Kraakman, R. (2002). "Property, contract, and verification: the *Numerus Clausus* problem and the divisibility of rights". Journal of Legal Studies 31, S373–S420.

Hardin, G. (1968). "The tragedy of the commons". Science 162, 1243–1248.

Harris, C., Vickers, J. (1985). "Patent races and the persistence of monopoly". Journal of Industrial Economics 33, 461–481.

Hart, O., Moore, J. (1990). "Property rights and the nature of the firm". Journal of Political Economy 98, 1119–1158.

Hayami, Y., Otsuka, K. (1993). The Economics of Contract Choice: An Agrarian Perspective. Oxford University Press, Oxford.

Hazlett, T.W. (1990). "The rationality of US regulation of the broadcast spectrum". Journal of Law and Economics 33, 133–175.

Hazlett, T.W. (1998). "Assigning property rights to radio spectrum users: why did FCC license auctions take 67 years?". Journal of Law and Economics 41, 529–575.

Hazlett, T.W., Munoz, R.E. (2004). "What really matters in spectrum allocation design", Manhattan Institute working paper.

Heller, M. (1998). "The tragedy of the anti-commons: property in transition from Marx to markets". Harvard Law Review 111, 621–688.

Heller, M., Eisenberg, R. (1998). "Can patents deter innovation? The anticommons in biomedical research". Science 280, 698–701.

Henderson, V., Ioannides, Y. (1983). "A model of housing tenure choice". American Economic Review 73, 98–113.

Hermalin, B. (1995). "An economic analysis of takings". Journal of Law, Economics & Organization 11, 64–86.

Hirsch, W. (1999). Law and Economics, 3rd edn. Academic Press, San Diego.

Hobbes, T. (1651) [1914]. Leviathon. Dent, London.

Holderness, C. (1989). "The assignment of rights, entry effects, and the allocation of resources". Journal of Legal Studies 18, 181–189.

Holmes, O.W. (1881) [1963]. The Common Law, DeWolfe, M. (Ed.). Little, Brown, Boston.

Holmes, O.W. (1897). "The path of the law". Harvard Law Review 10, 61–478.

Homstrom, B., Milgrom, P. (1991). "Multitask principal-agent analyses: incentives, contracts, asset ownership, and job design". Journal of Law, Economics, and Organization 7, 24–52.

Huffman, J. (1989). "A fish out of water: the public trust doctrine in a constitutional democracy". Environmental Law 19, 527–572.

Hughes, W., Turnbull, G. (1996). "Restrictive land covenants". Journal of Real Estate Finance and Economics 12, 9–21.

Hume, D. (1739–1740) [1978]. A Treatise on Human Nature. Clarendon, Oxford.

Hyde, W. (1981). "Timber harvesting in the Rockies". Land Economics 57, 630–637.

Innes, R., Polasky, S., Tschirhart, J. (1998). "Takings, compensation and endangered species protection on private lands". Journal of Economic Perspectives 12, 35–52.

Janczyk, J. (1977). “An economic analysis of the land title systems for transferring real property”. Journal of Legal Studies 6, 213–233.

Johnson, R.N., Gisser, M., Werner, M. (1981). “The definition of a surface water right and transferability”. Journal of Law and Economics 24, 273–288.

Kaplow, L. (1986). “The economics of legal transitions”. Harvard Law Review 99, 509–617.

Kaplow, L., Shavell, S. (1996). “Property rules versus liability rules”. Harvard Law Review 109, 713–790.

Kaplow, L., Shavell, S. (2002). “Economic analysis of law”. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 3. Elsevier, Amsterdam, pp. 1665–1784.

Karpoff, J. (1987). “Suboptimal controls in common resource management: the case of the fishery”. Journal of Political Economy 95, 179–194.

Keeton, W.P., Dobbs, D., Keeton, R., Owen, D. (1984). Prosser and Keeton on Torts, 5th edn. West Publishing Co., St. Paul.

Kitch, E. (1977). “The nature and function of the patent system”. Journal of Law and Economics 20, 265–290.

Klevorick, A. (1985). “On the economic theory of crime”. In: Pennock, J., Chapman, J. (Eds.), NOMOS XXVII, Criminal Justice. NYU Press, New York.

Knetsch, J., Borcherding, T. (1979). “Expropriation of private property and the basis for compensation”. University of Toronto Law Journal 29, 237–252.

Knight, F. (1924). “Some fallacies in the interpretation of social cost”. Quarterly Journal of Economics 38, 582–606.

Landes, W., Posner, R. (1987). “The Economic Structure of Tort Law”. Harvard University Press, Cambridge, MA.

Leiter, R. (1999). National Survey of State Laws, 3rd edn. Gale Research, Detroit.

Lerner, J. (1997). “An empirical exploration of a technological race”. RAND Journal of Economics 28, 228–247.

Leshy, J.D. (1987). The Mining Law. Resources for the Future, Washington.

Levmore, S. (2002). “Two stories about the evolution of property rights”. Journal of Legal Studies 31, S421–S452.

Libecap, G.D. (1978). “Economic variables and the development of the law: the case of Western mineral rights”. Journal of Economic History 38, 338–362.

Libecap, G.D. (1989). Contracting for Property Rights. Cambridge University Press, Cambridge.

Libecap, G.D., Smith, J. (2002). “The economic evolution of petroleum property rights in the United States”. Journal of Legal Studies 31, S589–S608.

Libecap, G.D., Wiggins, S. (1984). “Contractual responses to the common pool: prorationing of crude oil production”. American Economic Review 74, 87–98.

Libecap, G.D., Lueck, D. (2006). “Lowering transaction costs through institutional arrangements: the causes and consequences of the rectangular survey in assigning property rights to land”. National Science Foundation research project # 34-3416-00-0-79-340.

Locke, J. (1690) [1963]. Two Treatises of Government, P. Laslett (Ed.), 2nd edn. Cambridge University Press, Cambridge.

Lueck, D. (1989). “The economic nature of wildlife law”. Journal of Legal Studies 18, 291–323.

Lueck, D. (1994). “Common property as an egalitarian share contract”. Journal of Economic Behavior and Organization 25, 93–108.

Lueck, D. (1995a). “Property rights and the economic logic of wildlife institutions”. Natural Resources Journal 35, 625–670.

Lueck, D. (1995b). “The rule of first possession and the design of the law”. Journal of Law and Economics 38, 393–436.

Lueck, D. (1998). “Wildlife law”. In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 3. Stockton Press, New York, pp. 696–701.

Lueck, D. (2002). “The extermination and conservation of the American bison”. Journal of Legal Studies 31, S609–S652.

Lueck, D., Michael, J.A. (2003). “Preemptive habitat destruction under the Endangered Species Act”. Journal of Law and Economics 46, 27–61.

McFetridge, D., Smith, D. (1980). "Patents, prospects, and economic surplus: a comment". Journal of Law and Economics 23, 197–204.

McKay, B.J., Acheson, J.M. (1987). The Question of the Commons. University of Arizona Press, Tucson.

McMillan, J. (1994). "Selling spectrum rights". Journal of Economic Perspectives 8, 145–162.

McManus, J.C. (1975). "The costs of alternative economic organizations". Canadian Journal of Economics 8, 334–350.

Medina, B. (2003). "Augmenting the value of ownership by protecting it only partially: the 'Market-Overt' rule revisited". Journal of Law, Economics, and Organization 19, 343–372.

Merrill, T. (1985a). "Trespass, nuisance, and the cost of determining property rights". Journal of Legal Studies 14, 13–48.

Merrill, T. (1985b). "Property Rules, Liability Rules, and Adverse Possession". Northwestern University Law Review 79, 1122–1154.

Merrill, T. (1986). "The economics of public use". Cornell Law Review 72, 61–116.

Merrill, T. (1998). "Trespass and Nuisance". In: Newman, P. (Ed.), The New Palgrave Dictionary of Law and Economics, vol. 3. Stockton Press, New York, pp. 617–623.

Merrill, T.W., Smith, H.E. (2000). "Optimal standardization in the law of property: the *Numerus Clausus* principle". Yale Law Journal 110, 1–70.

Merrill, T.W., Smith, H.E. (2001). "What happened to property in law and economics?". Yale Law Journal 111, 357–398.

Miceli, T. (1991). "Compensation for the taking of land under eminent domain". Journal of Institutional and Theoretical Economics 147, 354–363.

Miceli, T. (1997). Economics of the Law: Torts, Contracts, Property, Litigation. Oxford University Press, New York.

Miceli, T., Segerson, K. (1996). Compensation for Regulatory Takings: An Economic Analysis with Applications. JAI Press, Greenwich.

Miceli, T., Segerson, K. (1994). "Regulatory Takings: When Should Compensation be Paid?". Journal of Legal Studies 23, 749–776.

Miceli, T., Sirmans, C.F. (1995a). "The economics of land transfer and title insurance". Journal of Real Estate Finance and Economics 10, 81–88.

Miceli, T., Sirmans, C.F. (1995b). "An economic theory of adverse possession". International Review of Law and Economics 15, 161–173.

Miceli, T., Sirmans, C.F. (1999). "The mistaken improver problem". Journal of Urban Economics 45, 143–155.

Miceli, T., Sirmans, C.F. (2000). "Partition of real estate; or, breaking up is (not) hard to do". Journal of Legal Studies 29, 783–796.

Miceli, T., Sirmans, C.F., Kieyah, J. (2001). "The demand for land title registration: theory with evidence from Kenya". American Law and Economics Review 3, 275–287.

Miceli, T., Sirmans, C.F., Turnbull, G. (2001). "The property-contract boundary: an economic analysis of leases". American Law and Economics Review 3, 165–185.

Miceli, T., Munneke, H.J., Sirmans, C.F., Turnbull, G.K. (2002). "Title systems and land values". Journal of Law and Economics 45, 565–582.

Michelman, F. (1967). "Property, utility, and fairness: comments on the ethical foundations of 'just compensation' law". Harvard Law Review 80, 1165–1258.

Mortensen, D. (1982). "Property rights and efficiency in mating, racing, and related games". American Economic Review 72, 968–979.

Munch, P. (1976). "An economic analysis of eminent domain". Journal of Political Economy 84, 473–497.

Munro, G.R., Scott, A.D. (1985). "The economics of fisheries management". In: Kneese, A., Sweeney, J. (Eds.), Handbook of Natural Resource and Energy Economics, vol. II. Elsevier, Amsterdam, pp. 623–676.

Nelson, R.H. (1977). Zoning and Property Rights. MIT Press, Cambridge, MA.

Nelson, R.H. (1995). Public Land and Private Rights. Island Press, Washington.

Netter, J., Hersch, P., Manson, W. (1986). "An economic analysis of adverse possession statutes". International Review of Law and Economics 6, 217–227.

Newberry, D., Stiglitz, J. (1979). "Sharecropping, risk-sharing, and the importance of imperfect information". In: Roumasset, J., Boussard, J., Singh, I. (Eds.), Risk, Uncertainty and Agricultural Development. University of California Press, Berkeley, pp. 311–339.

North, D.C. (1981). Structure and Change in Economic History. W.W. Norton, New York.

Nosal, E. (2001). "The taking of land: market value compensation should be paid". Journal of Public Economics 82, 431–443.

Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, New York.

Parisi, F., Schulz, N., Depoorter, B. (2006). "Duality in property: commons and anticommons". International Review of Law and Economics 25, 578–591.

Pigou, A.C. (1932). The Economics of Welfare. Cambridge University Press, Cambridge.

Pitchford, R., Snyder, C. (2003). "Coming to the nuisance: an economic analysis from an incomplete contracts perspective". Journal of Law, Economics, and Organization 19, 491–516.

Prendergast, C. (2000). "What trade-off of risk and incentives?". American Economic Review 90, 421–425.

Prendergast, C. (2002). "The tenuous trade-off between risk and incentives". Journal of Political Economy 110, 1071–1102.

Polinsky, A.M. (1980). "On the choice between property rules and liability rules". Economic Inquiry 18, 233–246.

Posner, R. (2003). Economic Analysis of Law, 6th edn. Aspen Law and Business, New York.

Rabin, E. (1984). "The revolution in residential landlord-tenant law: causes and consequences". Cornell Law Review 69, 517–584.

Riker, W.H., Sened, I. (1991). "A political theory of the origin of property rights: airport slots". American Journal of Political Science 35, 951–969.

Rose, C.M. (1986). "The comedy of the commons: custom, commerce, and inherently public property". University of Chicago Law Review 53, 711–781.

Rose, C.M. (1990). "Energy and efficiency in the realignment of common-law water rights". Journal of Legal Studies 19, 261–296.

Rose, C.M. (1998). "Evolution of property rights". In: Newman, P. (Ed.), The New Palgrave Dictionary of Law and Economics, vol. 2. Stockton Press, New York, pp. 93–98.

Rose-Ackerman, S. (1985). "Inalienability and the theory of property rights". Columbia Law Review 85, 931–969.

Rose-Ackerman, S. (1992). "Regulatory takings: policy analysis and democratic principles". In: Mercuro, N. (Ed.), Taking Property and Just Compensation. Kluwer Academic Publishers, Boston.

Rose-Ackerman, S. (1998). "Inalienability". In: Newman, P. (Ed.), The New Palgrave Dictionary of Law and Economics, vol. 2. Stockton Press, New York, pp. 268–273.

Sax, J. (1970). "The public trust doctrine in natural resources law". Michigan Law Review 68, 471–566.

Scott, A.D. (1955). "The fishery: the objectives of sole ownership". Journal of Political Economy 63, 116–124.

Sethi, R., Somanathan, E. (1996). "The evolution of social norms in common property resource use". American Economic Review 86, 766–788.

Schultz, T.W. (1968). "Institutions and the rising economic value of man". American Journal of Agricultural Economics 50, 1113–1122.

Shavell, S. (2004). Foundations of Economic Analysis of Law. Harvard University Press, Cambridge, MA.

Shick, B., Plotkin, I. (1978). Torrens in the United States: A Legal and Economic History and Analysis of American Land Registration Systems. Lexington Books, Lexington, MA.

Siegan, B. (1972). Land Use without Zoning. Lexington Books, Lexington, MA.

Smith A. (1776) [1909]. An Inquiry into the Nature and Causes of the Wealth of Nations. P.F. Collier, New York.

Smith, H.E. (2000). "Semicommon property rights and scattering in the open fields". Journal of Legal Studies 29, 131–169.

Smith, H.W. (2002). "Exclusion versus governance: two strategies for delineating property rights". Journal of Legal Studies 31, S453–S488.

Sprankling, J.G. (1996). "The anti-wilderness bias in American property law". University of Chicago Law Review 63, 516–590.

Stake, J. (1990). "Darwin, donations, and the illusion of the dead hand". Tulane Law Review 64, 705–781.

Stake, J. (1998a). "Inheritance law". In: Newman, P. (Ed.), The New Palgrave Dictionary of Law and Economics, vol. 2. Stockton Press, New York, pp. 311–321.

Stake, J. (1998b). "Land use doctrines". In: Newman, P. (Ed.), The New Palgrave Dictionary of Law and Economics, vol. 2. Stockton Press, New York, pp. 437–446.

Stake, J. (1999). "Decomposition of property rights". In: Bouckaert, B., DeGeest, G. (Eds.), Encyclopedia of Law and Economics, vol. II. Edward Elgar, Cheltenham, pp. 32–61.

Stake, J. (2001). "The uneasy case for adverse possession". Georgetown Law Journal 89, 2419–2474.

Stavins, R.N. (1995). "Transaction costs and tradeable permits". Journal of Environmental Economics and Management 29, 133–148.

Strange, W. (1995). "Information, holdouts, and land assembly". Journal of Urban Economics 38, 317–332.

Stevenson, G.G. (1991). Common Property Economics. Cambridge University Press, Cambridge.

Stigler, G.J. (1987). Memoirs of an Unregulated Economist. University of Chicago Press, Chicago.

Stroup, R., Baden, J. (1973). "Externality, property rights, and the management of our national forests". Journal of Law and Economics 14, 303–312.

Suen, W. (1989). "Rationing and rent dissipation in the presence of heterogeneous individuals". Journal of Political Economy 97, 1384–1394.

Sugden, R. (1986). The Economics of Rights, Cooperation and Welfare. Blackwell, London.

Ulen, T. (1992). "The public use of private property: a dual constraint theory of efficient government takings". In: Mercuro, N. (Ed.), Taking Property and Just Compensation: Law and Economics Perspectives of the Takings Issue. Kluwer Academic Publishers, Boston, pp. 163–198.

Umbeck, J. (1977). "The California gold rush: a study of emerging property rights". Exploration in Economic History 14, 197–206.

Wagner, T. (1995). "The enclosure of a common property resource: private and group ownership in equilibrium". Journal of Institutional and Theoretical Economics 151, 631–657.

White, M. (1975). "Fiscal zoning in fragmented metropolitan areas". In: Mills, E., Oates, W. (Eds.), Fiscal Zoning and Land Use Controls. Lexington Books, Lexington, MA.

White, M., Wittman, D. (1979). "Long-run versus short-run remedies for spatial externalities: liability rules, pollution taxes, and zoning". In: Rubinfeld, D. (Ed.), Essays on the Law and Economics of Local Governments. The Urban Institute, Washington, pp. 13–48.

Williams, S. (1983). "The requirement of beneficial use as a cause of waste in water resource development". Natural Resources Journal 23, 7–22.

Wittman, D. (1980). "First come, first served: an economic analysis of 'coming to the nuisance"'. Journal of Legal Studies 9, 557–568.

This page intentionally left blank

Chapter 4

LITIGATION

KATHRYN E. SPIER[*]

School of Law, Harvard University, and National Bureau of Economic Research

Contents

[*] The final draft of this manuscript was submitted in 2004. Because of the inevitable publication delays, this chapter cannot cover all of the important recent developments in the field. The author thanks A. Mitchell Polinsky and Steven Shavell for invaluable help writing this chapter and the Searle Fund at Northwestern University for financial support. Jim Dana, Andrew Daughety, Claude Fluet, Xinyu Hua, Bentley MacLeod, and Jennifer Reinganum. Kristina Steffenson McElheran contributed outstanding editorial assistance.

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)01004-3

Abstract

The purpose of this chapter is to survey the academic literature on the economics of litigation and to synthesize its main themes. The chapter begins by introducing the basic economic framework for studying litigation and out-of-court settlement. One set of issues addressed is positive (or descriptive) in nature. Under what conditions will someone decide to file suit? What determines how much is spent on a lawsuit? When do cases settle out of court? Important normative issues are also addressed. Are the litigation decisions made by private parties in the interest of society as a whole? Next, the chapter surveys some of the more active areas in the litigation literature. Topics include rules of evidence, loser-pays rules, appeals, contingent fees for attorneys, alternative dispute resolution, class actions, and plea bargaining.

Keywords

litigation, trials, courts, settlement, bargaining

JEL classification: C72, D82, K00, K41

1. Introduction

The topic of this chapter—Litigation—is one of the liveliest research areas in the field of Law and Economics. Why do some lawsuits go to trial while many others are resolved out of court? Is out-of-court settlement in the interest of society? How confident should a judge or jury be before finding for one party or the other? Should the losing party be required to reimburse the winning party's legal expenses? Should policy makers restrict the contracts between lawyers and their clients or are these contracts best left to the free market? To what extent should the ruling in a lawsuit be constrained by prior rulings in similar cases? The economic issues surrounding these and other important questions are surprisingly subtle and the techniques used to examine them have grown increasingly refined over the years. The purpose of this chapter is to survey the academic literature on the economics of litigation and to synthesize its main themes.[1]

In the United States, 2.3 million non-criminal cases were filed in federal courts and 20.1 million were filed in state courts in the year 2000.[2] These numbers, while staggering, underestimate the true "reach" of litigation because they exclude the countless cases that are never filed, yet are resolved in the shadow of litigation through private dispute resolution mechanisms. In all, legal services account for 1 percent of the United States' labor force and 1.3 percent of its gross domestic product.[3] The sheer magnitude of these numbers speaks to the importance of the study of litigation from both a positive and a normative perspective. Developing a better understanding of the economic forces behind litigation decisions is critical for the participants in litigation—plaintiffs, defendants, lawyers, and judges—and policy makers, alike.

The premise of this chapter is that the main purpose of the court system is to facilitate value-creating activities and deter value-destroying activities through the enforcement of contracts and laws.[4] Mortgage contracts, for example, create economic value by allowing would-be homeowners to borrow against their future earnings. If the court system refused to uphold these contracts (preventing the bank from foreclosing in the event of non-payment, for example) then borrowers would default on their loans more often and, in anticipation, banks would choose to lend less money. Similarly, insurance

[1] Previous surveys include Cooter and Rubinfeld (1989), Hay and Spier (1998), and Daughety (2000).

[2] Ostrom, Kauder, and LaFountain (2001, p. 13). The state court figure also does not include traffic cases and juvenile cases.

[3] Statistical Abstract of the United States (2001, tables 596 and 641, pp. 384 and 418).

[4] The court system has other purposes, of course, some of which are also economic in nature. For instance, the courts can also provide a mechanism for *creating rules*. In particular, judges are often put in the position of "filling gaps" when private contracts are incomplete—in effect, creating a rule where one did not previously exist. In common law regimes, judges have the power to change or modify inappropriate or obsolete laws, decisions that may bind on future decisions through precedent. Moreover, litigation is a mechanism for *disseminating information*—it communicates to would-be offenders, for example, the consequences of their actions. Observing litigation may also teach us right from wrong and inform us of the risks associated with various activities.

policies (e.g., car insurance, homeowners insurance, life insurance, and health insurance) create economic value because risk-averse individuals would rather pay for their expected losses upfront through installments or insurance premiums. The value that the consumers derive from avoiding risk would be forgone if the insurance contracts were not upheld by the court system. Criminal penalties create value by deterring would-be criminals from stealing or destroying the valuable property of others. If these penalties were unenforceable by the court system, then deterrence would be compromised and more crimes would be committed.[5]

In an ideal world, the court system would be accurate, unbiased, and free. The enforcement of rules would take place immediately and no transactions costs would be incurred. But the world is far from ideal. Mistakes are made at trial: acquitted defendants may in fact be guilty of the charges and convicted defendants may be innocent. Judges and juries may also bring their personal or political biases into the courtroom: a sympathetic jury may award astronomical damages to a severely injured child or be overly harsh towards a corporate executive with deep pockets. Errors, biases, and expected litigation costs also distort the economic activities that take place in the shadow of litigation. A physician, for example, may prescribe unnecessary diagnostic tests for patients and may avoid certain practice areas (such as obstetrics) because of these added litigation risks.

The chapter is organized as follows. Section 2 introduces the basic economic framework for studying litigation and out-of-court settlement. Some of the issues are positive (or descriptive) in nature. For instance, under what conditions will someone decide to file suit? What determines how much is spent on a lawsuit? When do cases settle out of court? This section also addresses important normative issues. In particular, it explores whether the litigation decisions made by private parties are in the interest of society more broadly. Section 3 then uses the basic framework to survey some of the more active areas in the litigation literature. Topics include the rules of evidence, loser-pays rules, appeals, contingent fees for attorneys, alternative dispute resolution, class actions, and plea bargaining. Concluding remarks follow.

2. Basic framework

This section presents the basic economic framework used in the study of litigation. Although future sections of this chapter consider more sophisticated models, this section assumes that there is a single case involving two *litigants*: one *plaintiff* and one *defendant*. The plaintiff is the injured party who seeks compensation for her injuries. The defendant is the party who is potentially responsible for the plaintiff's injuries. We assume that litigation is costly for both the plaintiff and the defendant. Indeed, the average

[5] Jail time can also create value by incapacitating the criminal, thereby preventing future crimes.

hourly billing rate of a law firm partner in the United States in 2001 was $246.[6] Because of the private expenses associated with using the court system to resolve disputes, plaintiffs and defendants (and arguably the lawyers who represent them) have economic incentives to weigh their options carefully and make prudent decisions in litigation.

Section 2.1 abstracts from settlement decisions and considers the plaintiff's decision to pursue litigation. Section 2.2 considers the incentives of the litigants to resolve their case out of court through private settlement.

2.1. *The decision to litigate*

2.1.1. *Bringing suit*

A plaintiff's decision to litigate hinges on her private benefits from pursuing the case as well as her private costs of pursuit. [7] She will rationally choose to bring suit when the expected gross return from litigation, call it x, exceeds the cost of bringing the case to trial, c_p. The gross return, x, could reflect the expected judgment at the end of a long and costly trial or a settlement that takes place at some time prior to the trial. More generally, it could reflect issues that are somewhat beyond the scope of the current dispute, such as the impact that a court decision will have on future cases or the plaintiff's concern for her business reputation. The plaintiff's litigation cost, c_p, would typically include money paid to private attorneys, but it could also include the plaintiff's personal cost of effort, time, and any other opportunity costs associated with the plaintiff's involvement in the lawsuit. As with other economic decisions, the plaintiff will choose to pursue litigation when the strategy has positive expected value, $x > c_p$, and she will choose to not pursue litigation if the case has a negative expected return, $x < c_p$.[8]

2.1.2. *Private litigation spending*

In practice, the plaintiff—often with the help of her attorney—must decide how much time and effort to invest in the lawsuit. In other words, the plaintiff's private litigation costs are endogenous rather than exogenous. The plaintiff's investment choice will reflect both the underlying facts of the case and the beliefs that the plaintiff holds about the future of the case—including those concerning the investments and responses of the defendant.

[6] The 2001 Survey of Law Firm Economics (pp. 11–39).

[7] In this subsection I will not distinguish between cases that are settled and those that go all the way to final judgment. Settlement will be taken up in the next subsection, and appeals are considered in section 3.

[8] In this section, it is assumed that each litigant bears his or her own litigation costs, regardless of the outcome at trial. This known as the American Rule. Under the English Rule, on the other hand, the loser in litigation is required to reimburse the winner for certain expenses. The economic effects of these rules are discussed in section 3.

Formally, suppose that the plaintiff's expected recovery from litigation depends on her own spending as well as that of the defendant. The litigation investments made by the plaintiff and defendant, c_p and c_d, respectively, affect the plaintiff's future recovery at trial, $x(c_p, c_d)$. This function is increasing in the first argument and decreasing in the second.[9] The equilibrium investments can be viewed as the outcome of a non-cooperative game (which could be simultaneous or sequential) where the plaintiff maximizes her expected litigation return,

$$x(c_p, c_d) - c_p,$$

and the defendant minimizes his expected total payments,

$$x(c_p, c_d) + c_d.$$

As described earlier, the plaintiff will bring suit when her anticipated return from litigation, $x(c_p, c_d) - c_p$, is positive.

The structure of this game is similar to those of other types of contests, including patent races, tournaments in internal labor markets, and a variety of sports. The dynamic strategies employed by the participants in these contests hinge on the anticipated reactions of the rival—each player would like his opponent to "back off" and invest less in the contest (see Dixit, 1987; Katz, 1988).[10] To this end, the plaintiff might derive a strategic benefit from aggressive spending, for example, when the defendant's best-response function is decreasing in the plaintiff's investment (i.e., is downward-sloping). Conversely, the plaintiff would benefit from a commitment to lower spending levels if the defendant's best-response curve slopes upward. Either situation can arise in the general framework.[11] Along these lines, a credible commitment by the defendant to build a defensive war chest could lower the plaintiff's expected return from litigation, leading the plaintiff to drop the claim altogether.

2.1.3. *Normative implications*

The private decisions of the plaintiff and defendant to invest time and money in a lawsuit are not generally aligned with the interests of society as a whole. This section highlights why the plaintiff's private incentive to use the legal system diverges from the social

[9] This general structure captures the specific case where private litigation expenditures influence the plaintiff's probability of winning but the damages conditional on winning are fixed.

[10] See Bulow et al. (1985) for a general discussion of strategic commitment in competitive environments.

[11] Interestingly, the mathematical condition for the defendant's best-response curve to slope down—that the cross partial derivative of the recovery function is positive—implies that the plaintiff's best-response curve is upward sloping. Although the plaintiff may want to commit to aggressive spending, the defendant might actually refrain from tough spending strategies since aggressive spending would only elicit an aggressive response from the plaintiff. For some functional specifications, the best-response curves are backward-bending, i.e., increasing in some range and decreasing in others. This, of course, raises challenges for both theoretical inquiry and empirical estimation.

incentive.[12] In some circumstances, the plaintiff will not litigate often enough (or spend too little on the case), while in others she will litigate too often (or spend too much on the case).

To better illustrate both situations, first consider a simple moral hazard example where the potential defendant can take precautions (which require effort and/or monetary investments) to reduce the probability of an accident. If the defendant fails to take precautions, then the accident occurs with probability p_0 and the plaintiff sustains damages x. By investing e, the defendant reduces the probability of harm to $p_1 < p_0$. The defendant's decision to take precautions will hinge, of course, on what he expects the plaintiff to do following the accident. If the plaintiff's litigation costs are high relative to the magnitude of harm, $c_p > x$, then the plaintiff will choose not to litigate following an accident. Anticipating no future legal action, the defendant would have no incentive to take precautions *ex ante*. Conversely, if the plaintiff's litigation costs are relatively low, $c_p < x$, then the plaintiff will certainly litigate in the event of an accident. In this case, the defendant will rationally choose take precautions if his total expected cost associated with taking precautions, $e + p_1(x + c_d)$, is smaller than his total expected cost associated with being careless, $p_0(x + c_d)$.[13]

Insufficient litigation activity It is easy to construct examples where the plaintiff has an insufficient incentive to bring suit. Take, for example, the case where $p_1 = 0$ and $p_0 = 1$. Accidents are completely avoided if the defendant takes precautions and occur with certainty, otherwise. Suppose further that $e < x$, so that it is socially efficient for the defendant to take precautions. If the plaintiff could commit ahead of time to aggressively pursue litigation in the event of an accident, then, in anticipation, the defendant would rationally choose to take precautions to avoid the accident to begin with. There would be no accidents and no litigation costs spent—the first-best outcome! But the plaintiff's commitment to litigate following an accident may not be credible: when the plaintiff's cost of pursuing a case is high relative to the expected judgment, i.e., $c_p > x$, the plaintiff will choose not to pursue the claim. In anticipation, the defendant has no incentive to take precautions and deterrence is compromised.

This problem may be especially severe if the defendant has deep pockets and a reputation for tenacious defense spending. In the shadow of a powerful defendant, an injured plaintiff may rationally decide *ex post* to scale back or drop the suit altogether. This problem may be especially severe in situations where the defendant's failure to take precautions affects many victims rather than just one. Oil spills (e.g., the Exxon Valdez incident) and poorly-designed consumer products (e.g., the Ford Pinto) are two examples where the harms were broadly dispersed among victims while the responsibility for harm was concentrated. In situations like these, an individual victim's decision to sue

[12] See Shavell (1982b, 1997) for more thorough—and excellent—discussions of these important issues.

[13] The defendant's total expected cost includes the cost of precautions, the future expected liability, and the future expected litigation costs.

is a public good: the indirect effect on deterrence helps the population of victims more broadly. Due to the incentive of victims to free ride on the litigation efforts of a few (who must incur the litigation costs) there is likely to be insufficient litigation, from a societal perspective.

These ideas are also important in criminal law enforcement. When a victim of a crime presses charges, the benefits of a successful prosecution—which may include the incapacitation of dangerous individuals in addition to deterrence—will accrue to society more broadly. Furthermore, a criminal conviction does not entitle the victim to any monetary compensation (although the victim can sometimes successfully pursue a civil action.) The victim's costs of pressing charges may be significant, including the time and effort associated with the collection of evidence and appearing as a witness as well as the psychological costs associated with recounting bad experiences. There may also be indirect costs associated with embarrassment or loss of reputation or social standing as the victim of a crime (e.g., the stigma associated with rape that makes survivors reluctant to testify). In these situations, the socially desired level of litigation exceeds what would otherwise be observed in equilibrium.

Excessive litigation activity Using the simple model outlined at the beginning of this section, it is not difficult to construct examples where the plaintiff's private incentive to bring suit is *socially excessive*. Take the stark example where the potential defendant's precautions are totally ineffective: accidents occur with the same likelihood whether or not the defendant took care, $p_1 = p_0$. Note that the defendant would never take precautions in this extreme case. (Indeed, it would be socially wasteful if he did.) The plaintiff will sue the defendant following an accident when litigation has positive expected value, $c_p < x$, and the defendant will be forced to bear the litigation costs as well. This is socially wasteful. There is nothing the defendant could have done to avoid the accident, and consequently the litigation costs are a pure deadweight loss for society.

This type of divergence between private incentives and social objectives is very important in practice. Take, for example, the case of a divorcing couple battling in court over the financial assets acquired during their marriage and perhaps the custody of their children. One can imagine situations where every dollar spent by the wife is exactly cancelled out by a dollar spent by the husband. Both parties may end up spending a lot of money, time, and effort just to stand still! While divorce lawyers could profit handsomely in these types of cases, the litigants themselves and their children sometimes end up poor (not too mention bitter and emotionally drained).

This type of divergence can also arise in situations where precautions taken by the potential defendant reduce the expected social harm. Let's return to the more general case where $p_1 < p_0 < 1$. Suppose the plaintiff has a private incentive to litigate following an accident and that the defendant, in anticipation, takes costly precautions to reduce the likelihood of an accident (and, consequently, his future liability). The total social cost in this scenario is $e + p_1(x + c_p + c_d)$, the sum of the defendant's cost of precautions, the expected damages suffered by the plaintiff, and the expected litigation costs of both parties. If the plaintiff refrained from pursuing litigation (and the defendant rationally refrained from taking precautions) then the total social cost would simply be p_0x. This latter case may be preferable. In particular, if $(p_0-p_1)x < p_1(c_p+c_d)+e$ then the level of private litigation activity is too high from the point of view of society as a whole. Intuitively, the society-wide benefit of litigation—the lower accident rate that results when the defendant takes more care—is outweighed by the total costs of litigation.

Discussion If the direction of the divergence between private and social incentives is clear, then a variety of public policies may be introduced to bring them into greater alignment. Damage multipliers, where a plaintiff can receive several times her actual damages, or punitive damages can give individual plaintiffs an additional incentive to litigate claims that would otherwise have negative expected value.[14] The ability of individual plaintiffs to consolidate their claims—through a joinder, perhaps—and the ability of lawyers to represent the interests of dispersed plaintiffs through class representation mitigates the externalities that otherwise would constitute an obstacle to recovery and deterrence. Conversely, other policies can be used to reduce the level of litigation. Strict liability, where a defendant is liable for a plaintiff's damages regardless of his level of care, can lead to high levels of litigation activity. The negligence rule, where the defendant is liable only if he took inadequate care, can give the defendant the appropriate incentives for care while reducing the total volume of litigation and its corresponding social costs. These brief examples are given for illustrative purposes only—the subtleties of these research areas and many others will be discussed in greater detail in section 3.

2.2. *Out-of-court settlement*

The preceding discussion focused on the decision of a plaintiff to bring a lawsuit while abstracting from the private settlement of litigation. We will now focus on the private settlement of legal claims. In practice, the vast majority of cases that are filed ultimately settle before trial and countless others are settled before a case is filed at all. Less than 4 percent of civil cases that are filed in the U.S. State Courts go to trial.[15] In the U.S. Federal Courts, only about 2 percent of civil cases go to trial.[16] The fact that most private litigants choose to "opt out" of formal litigation channels should not be too surprising. The pursuit of litigation is expensive, time-consuming, and distracting. In short, trials are a decidedly inefficient way for private parties to resolve their disputes.

As above, suppose that x is the expected judgment at trial and let c_p and c_d be the plaintiff and defendant's trial costs, respectively. If the case went to trial, the plaintiff would receive a net payoff of $x - c_p$, the expected judgment minus his litigation costs,

[14] Polinsky and Rubinfeld (1988b) argue that compensatory damages, where the plaintiff's award is exactly equal to her damages, is not generally optimal. The optimal damages would either be adjusted upward or downward, depending upon the circumstances.

[15] Ostrom, Kauder, and LaFountain (2001, p. 29).

[16] Judicial Business of the United States Courts (2001, p. 154, table C-4).

and the defendant's payoff would be $-x - c_d$. Although x represents a simple transfer from the defendant to the plaintiff, the total cost of litigation, $c_p + c_d$, is a deadweight loss—"money down the drain," so to speak. The plaintiff and the defendant can typically avoid this loss through a private agreement to end the dispute before the litigation costs are incurred. Specifically, a binding settlement contract that specifies a transfer from the defendant to the plaintiff, $S \in (x - c_p, x + c_d)$, leaves both the plaintiff and the defendant better off than they would be from going to trial.

The insight that out-of-court settlement Pareto dominates trials (at least for the litigants themselves) raises a number of interesting questions. For what amount will the case settle? Will the defendant agree to settle if the plaintiff has a negative expected value claim, $x - c_p < 0$? When will the case settle—shortly after filing or on the courthouse steps? Why do some cases fail to settle? Is private settlement in the interest of society, more broadly? We will see that the answers depend on many factors, including the timing of offers and counteroffers, the information and beliefs of the two litigants, and the way that the particular lawsuit fits into the broader economic, legal, and strategic environment.

2.2.1. Settlement with symmetric information

The simplest economic framework considers settlement under symmetric information. Suppose that the two litigants have exactly the same beliefs about what will happen if the case goes to trial—i.e., they are symmetrically informed about the stakes of the case, the litigation costs, and all other relevant parameters. We will begin by assuming that $x - c_p > 0$, so the plaintiff clearly has a credible threat to take the case all the way to trial. Later, we will relax this assumption and assume that $x - c_p < 0$. Even though, in this latter instance, the case has negative expected value if pursued all the way to trial, the plaintiff may nevertheless succeed in extracting a settlement in equilibrium.

"Lumpy" litigation costs Consider a simple extensive form game where there are $T - 1$ rounds of bargaining before a costly trial in round T. The litigants alternate in making offers. Suppose that the plaintiff is designated to make the last settlement offer before the trial, call it S_{T-1}. If the case fails to settle then x is transferred from the defendant to the plaintiff and the litigation costs, c_p and c_d, are incurred. The costs of litigation are lumpy in the sense that they are all incurred at trial instead of gradually over time. Finally, suppose that the two litigants discount time at the same rate and let δ be their common discount factor.

This game is solved by backwards induction. In period $T - 1$ the defendant will accept any offer that is better than going to trial, where he would pay $x + c_d$ in total. The plaintiff's best offer in period $T - 1$ is $S_{T-1} = \delta(x + c_d)$, minus a penny perhaps.[17]

[17] For notational simplicity, I will assume that the defendant will accept an offer when he is indifferent between accepting and rejecting. The hypothetical penny would make him strictly prefer acceptance, but it would muddy the algebra.

Working backwards, we consider period $T - 2$, where the defendant has the right to make a settlement offer. In this period, both litigants realize that if the case fails to settle in period $T - 2$ then it will certainly settle for $S_{T-1} = \delta(x + c_d)$ in the next round. The very least that the plaintiff is willing to accept is therefore δS_{T-1}, and so the defendant would offer no more than $S_{T-2} = \delta^2(x + c_d)$ and the plaintiff would accept. This logic suggests that the case will settle in the first round for $S_1 = \delta^{T-1}(x + c_d)$.

The allocation of the bargaining surplus is sensitive to the timing of the settlement offers. Suppose instead that the defendant is the one to make the last offer. In this case, the defendant would offer $S_{T-1} = \delta(x - c_p)$ in the last round and the plaintiff would accept. Working backwards to the first offer we would find $S_1 = \delta^{T-1}(x - c_p)$. Comparing these two cases—one where the defendant makes the final offer and one where the plaintiff makes the final offer—delivers an important insight: the party who makes the last offer succeeds in extracting all of the bargaining surplus! In this alternating offer game with all litigation costs being incurred in a "lump" at trial, being last is much more important than being first.

The bargaining surplus would, of course, be more evenly shared under different assumptions about the timing of offers and counteroffers. Suppose instead that the two litigants flip a coin in each bargaining round to determine who will make the offer. In the last round (if it were reached) the parties would settle, on average, for $S_{T-1} = \delta[x + (c_d/2) - (c_p/2)]$. Working backwards, one can see that the case could settle in the first round for $S_1 = \delta^{T-1}[x + (c_d/2) - (c_p/2)]$, regardless of who makes the offer. Note that if the plaintiff's and defendant's litigation costs are of roughly the same magnitude, then the settlement amount would accurately reflect the discounted expected judgment at trial.

It is interesting to note in this example that, although it is efficient for the case to settle before trial, *it is no more efficient for the case to settle in round 1 than in round* $T - 1$. In the random-offer framework of the previous paragraph, if the case settles in round 1 the plaintiff's payoff is $S_1 = \delta^{T-1}[x + (c_d/2) - (c_p/2)]$. If settlement is delayed so that the case settles on the courthouse steps instead, then the plaintiff's payoff (discounted back to round 1) is the same: $\delta^{T-1}S_{T-1} = \delta^{T-1}[x + (c_d/2) - (c_p/2)]$. To put it somewhat differently, the plaintiff is indifferent between settling early and settling late and a similar logic shows that the defendant is indifferent as well. The reason for this is simple: there is no inefficiency associated with delay when the costs of litigation are "lumpy" and are all borne at trial.

Finally, the simple framework presented here differs in several important respects from the related (and certainly more famous) framework of bilateral trade. Suppose that a buyer's valuation for a good or service, B, exceeds the seller's cost, C. Following trade, the buyer enjoys surplus $B - P$ and the seller enjoys profit $P - C$. It is clearly in the buyer's and seller's mutual interest for trade to take place sooner rather than later because discounting causes the total value that must be shared between them to diminish. If they trade immediately, their joint surplus is $B - C$. If they were to wait and trade in period $T - 1$ (say) then their joint surplus would be $\delta^{T-1}(B - C)$. In bilateral trade, discounting causes the total "pie" to shrink. This is not the case in the settlement framework. When all of the costs of litigation are incurred at trial, the bargaining surplus in each bargaining round is zero: the plaintiff receives exactly what the defendant pays.

Divisible litigation costs The previous section assumed that all of the costs of litigation were incurred at trial. In reality, the costs of litigation are incurred by the litigants in a variety of ways while preparing for the trial. In addition to the direct costs of trial preparation, there may be costs of distraction (as litigants focus on lawsuits rather than their jobs and families) and risk management issues associated with open claims.[18] One can easily extend the earlier framework by assuming that the costs of litigation are divided among the T rounds—pretrial bargaining rounds as well as the trial round. In contrast to the scenario with lumpy litigation costs, there is a unique subgame perfect equilibrium where the case settles in round 1 (Bebchuk, 1996).[19] When litigation costs are incurred over time, then delay is inefficient and there are strong economic incentives to settle early.

Negative expected value (NEV) claims The previous sections assumed that the plaintiff had a credible threat to take the case to trial if it did not settle. We will now suppose that the plaintiff has a "negative expected value claim," one that would be unprofitable for the plaintiff if pursued all the way to trial. In particular, we assume $x - c_p < 0$. The defendant will certainly refuse to pay a cent if the plaintiff cannot credibly commit to litigation. Can the plaintiff succeed in extracting a settlement offer from the defendant under these circumstances?

The key to credibility for the plaintiff hinges on some of the factors identified above: (i) the divisibility of the litigation costs and (ii) the timing of offers. Suppose that the costs are "lumpy" as above—all of the litigation costs are borne at trial—and that the plaintiff has the option of dropping the case at any point in time. Subgame perfection implies that if settlement negotiations fail, then the plaintiff will surely drop the case on the courthouse steps. Backwards induction implies that the defendant would never make or accept an offer to settle.

It would be a mistake to conclude that negative expected value claims cannot succeed, however. Bebchuk (1996) shows that when the litigation costs are divisible and spread over the bargaining phase, then the set of circumstances under which a plaintiff can succeed in extracting a settlement is broader. To see why, suppose for simplicity that the litigation costs, c_p and c_d, are equally divided among the T periods and that there is no discounting of time ($\delta = 1$). Although the case may begin with negative expected value, $x - c_p < 0$, it can be "transformed" into a positive expected value case when T, the number of periods, is sufficiently large. To see why, imagine that the case actually reaches the courthouse steps. Most of the litigation costs are sunk at that point in time.

[18] A corporate defendant, for example, may need to keep additional resources on hand in anticipation of the possibility of a large adverse judgment or other surprises.

[19] With discounting, this expression would be modified to reflect the fact that the costs are not all borne at the same time.

When standing on the courthouse steps, the plaintiff has a credible threat to proceed to trial when $x - c_p/T > 0$. (The plaintiff's ultimate threat to litigate is stronger when T, which captures the divisibility of the costs, is larger.) If they were to "flip a coin" at that point, the case would settle on average for $S_{T-1} = x + (c_d/2T) - (c_p/2T)$. Working backwards[20] (and assuming that the plaintiff's litigation costs are not too large relative to the defendant's), we would find that the case could settle for approximately $S_1 = x + (c_d/2) - (c_p/2)$ in round 1. The plaintiff's threat to litigate is thus credible at each stage of the game!

2.2.2. Settlement with asymmetric information

In the simplest settlement games with symmetric information—including the examples discussed above—cases either settle out of court for a positive amount or are dropped without future costs being incurred. In short, settlement with symmetric information is privately efficient. In this section we will see that bargaining may be privately inefficient if the litigants are asymmetrically informed. These inefficiencies result in costly trials as well as costly delays in negotiations.

Asymmetric information has a variety of sources in litigation and manifests itself in a variety of ways. The plaintiff, for example, may have first-hand knowledge of the level of damages she has suffered; the defendant may have first-hand knowledge about his degree of involvement in (or liability for) the accident. Both litigants may know better the credibility of their own witnesses and the quality and work ethics of their lawyers.[21] It is important to note that some of this information will become commonly known over time—the parties may learn a great deal through pretrial settlement proceedings, for example. Other information may not come to light at all, but can nevertheless affect the trial outcome. The salient point here is that the revelation of information—through pretrial discovery activities and through formal legal proceedings—is both privately and socially costly. All else equal, litigants have private incentives to settle their dispute before the costs are sunk.

For the rest of this section we will assume that the defendant has private information about x, the expected judgment at trial. A similar analysis would follow if it were the plaintiff who had the private information, instead.[22] Formally, he observes the parameter x where x is drawn from a probability density function $f(x)$ on $[\underline{x}, \bar{x}]$ with cumulative

[20] The second to last offer would be $S_{T-2} = x + (c_d/T) - (c_p/T)$ on average, and so on.

[21] The sources of asymmetric information mentioned in this paragraph all affect the expected judgment at trial. The parties may also be privately informed about their own level of risk aversion or their discount rates, as in Farmer and Pecorino (1994). Private information along these lines will certainly affect the settlement offers that the information holder is willing to make or receive, but is not directly payoff relevant to the rival. Private information could also be endogenously determined by the litigants' effort choices, as in Hay (1995).

[22] Schweizer (1989) and Daughety and Reinganum (1994) consider extensive form games with two-sided asymmetric information. Spier (1994b), discussed in more detail later, explores two-sided asymmetric information settings with mechanism-design techniques.

density $F(x)$. For reasons that will become apparent later, we will also assume that this distribution has a *monotone hazard rate*, so $[1 - F(x)]/f(x)$ is everywhere decreasing in x.

Screening models Starting with P'ng (1983) and Bebchuk (1984), many papers have considered a somewhat special framework where the uninformed player—the plaintiff in our example—makes a single take-it-or-leave-it settlement offer, S, before a costly trial. The settlement offer "screens" the defendants into two groups: those who accept and those who reject. When faced with a settlement offer, S, the defendant will certainly choose to accept the offer if it is lower than what he would expect to pay at trial, $S < \delta(x + c_d)$. The settlement offer corresponds with a cutoff, $\hat{x} = \delta^{-1}S - c_d$, where defendants with types above the cutoff accept the plaintiff's offer to settle and those below the cutoff go to trial instead. Note that the sample of cases that go to trial is not a random sample—a defendant who is more confident about his prospects at trial (a defendant with low x) is more likely to reject the settlement offer and go to trial. Thus, the cases that go to trial would, on average, have lower judgments than those that settle out of court.

This case-selection result would, of course, be reversed if the plaintiff had private information instead and the defendant could make a take-it-or-leave-it offer before trial. The defendant's settlement offer would correspond to a cutoff where plaintiffs with low x's accept the offer and those with high x's reject the offer and go to trail. In this instance, cases that go to trial would, on average, have higher judgments than those that settle. This observation should make it clear to the reader that the empirically testable implications of these models depend very strongly on the source of the information asymmetry.

For now, we return to the case of private information on the part of the defendant. If given the power to make a take-it-or-leave-it offer, the plaintiff would choose her settlement offer to maximize her expected profit. Since each settlement offer corresponds to a unique cutoff, we may write the plaintiff's optimization problem as a function of the cutoff, $\hat{x}$:

$$\underset{\hat{x}}{Max} \int_{\underline{x}}^{\hat{x}} \delta(x - c_p) f(x) dx + [1 - F(\hat{x})]\delta(\hat{x} + c_d).$$

The first term represents the payments associated with the confident defendant types who reject the settlement offer and go to trial. Note that we are assuming that the plaintiff has to bear her own litigation costs in these cases. Fee-shifting rules, including the English rule, are discussed in section 3. The second term reflects payments, $S = \delta(\hat{x} + c_d)$, from the defendant types above the cutoff, $\hat{x}$. An interior solution, if it exists, is characterized by the following first-order condition:[23]

$$1 - F(\hat{x}) - (c_p + c_d) f(\hat{x}) = 0.$$

[23] The monotone hazard rate assumption guarantees that this solution is unique.

This condition may be understood intuitively: when the plaintiff raises her offer slightly from $S = \delta(\hat{x} + c_d)$ to $S = \delta(\hat{x} + c_d + \Delta)$, there are both benefits and costs. The benefit is that those defendants with types above $\hat{x} + \Delta$ will pay $\hat{x} + \Delta$ more than before, a benefit which is approximately $\Delta\delta[1 - F(\hat{x})]$, the discounted additional payment, $\Delta\delta$, multiplied by the probability of acceptance $1 - F(\hat{x})$. The cost is that defendants with types between $\hat{x}$ and $\hat{x} + \Delta$ go to trial instead of settling. This cost to the plaintiff is approximately $\Delta\delta(c_p + c_d)f(\hat{x})$. This includes both the plaintiff's cost of litigation (borne directly by the plaintiff) and the defendant's cost of litigation (borne indirectly by the plaintiff through the foregone settlement offer).

The first order condition implies that at least some cases will settle—the plaintiff will certainly make a settlement offer that is accepted by the most liable defendant. This can be seen by setting $\hat{x} = \bar{x}$. The left hand side of the first-order condition is negative when evaluated at this value, implying $\hat{x} < \bar{x}$. Furthermore, if $(c_p + c_d) > [1 - F(\underline{x})]/f(\underline{x})$ then the plaintiff will make a "low ball" offer, $S = \delta(\underline{x} + c_d)$, that all defendant types will accept. Intuitively, all cases settle if the litigation costs outweigh the degree of asymmetric information. An interior solution exists when the costs of litigation are not too high, i.e., when $(c_p + c_d) < [1 - F(\underline{x})]/f(\underline{x})$.

This screening example assumed that the plaintiff had a credible commitment to pursue the case all the way to trial. This is not necessarily true. The defendant types who reject the offer are the ones who believe that they have strong cases (i.e., they have low x's). The plaintiff, understanding this, may therefore have an incentive to drop the case following the rejection of her settlement offer! This credibility issue may be seen most readily by considering the case where $\underline{x} = 0$ and the litigation costs are large. The preceding analysis tells us that the plaintiff should offer to settle for $S = \delta(\underline{x} + c_d) = \delta c_d$. But if the defendant rejects the offer, the plaintiff may rationally believe that the defendant is a low type and will choose to drop the claim. If so, the defendant has an incentive to reject the settlement offer. Nalebuff (1987) extends this framework to incorporate a credibility constraint and shows that when the constraint is binding, the equilibrium settlement offer is higher than before. (A higher offer implies that the average return at trial from the bottom of the truncated distribution is higher as well, making the plaintiff's commitment to proceed credible.)

The screening example above also assumed that there was a single offer before trial. Spier (1992a) extends Bebchuk's framework to consider a sequence of settlement offers before trial. When litigation costs are "lumpy," a striking pattern emerges: the plaintiff waits until the very last moment to offer $S_{T-1} = \delta(\hat{x} + c_d)$, where $\hat{x}$ is defined above. That is, all settlement occurs on the courthouse steps! This result is important for several reasons. First, we often observe these 11th hour settlements in practice, demonstrating the practical relevance of the asymmetric information framework. Second, this result implies that the common way of modeling these bargaining games—that of take-it-or-leave-it offers—is less restrictive than it may first appear. The finitely-repeated screening model where all of the costs are borne at trial is equivalent to the simple model with a single offer.

The reason why the plaintiff refrains from early offers is not hard to see. Suppose that, for the sake of argument, the plaintiff could commit to a take-it-or-leave-it offer in round 1, taking the defendant all the way to trial if he rejects the round 1 offer. The best offer that she could possibly make is $S_1 = \delta^{T-1}(\hat{x} + c_d)$, implementing the very same cutoff as above. Note that the plaintiff is indifferent between receiving $S_1 = \delta^{T-1}(\hat{x}+c_d)$ in round 1 and the present discounted value of receiving $S_{T-1} = \delta(\hat{x}+c_d)$ on the courthouse steps. As discussed earlier, when the plaintiff and defendant discount time at the very same rate, then the passage of time alone does not impose costs on the two litigants. It is not credible, however, for the plaintiff to make a take-it-or-leave-it offer in round 1. Following a rejection, the plaintiff would update her beliefs about the remaining distribution and make a lower settlement offer. In anticipation, the defendant would wait to receive a better offer. The plaintiff can thus use delay to her advantage. By postponing settlement talks until the last moment, the plaintiff optimally extracts rents from the defendant.

Spier (1992a) also shows that when the costs of litigation are divisible over time, the plaintiff's optimal strategy does involve some settlement in each round. In particular, there is more settlement in the first rounds than in the middle and, if the costs borne at trial are disproportionately large, one ought to observe a pronounced deadline effect where many cases settle on the courthouse steps (generating a "U-shaped" pattern of settlement, overall). Importantly, this result is still obtained in the limit as the number of rounds, T, approaches infinity. The reason that this result differs from the familiar Coase Conjecture is that the passage of time before settlement does not screen among defendant types in the ordinary sense. Since they discount time at the same rate, the different defendant types have the same preferences between a settlement offer in round 1 and a settlement offer in round T. The types differ, however, in their preference for going to trial, generating delay in equilibrium. See also the empirical work of Fournier and Zuehlke (1996).

Signaling models We will now suppose that the informed defendant, rather than the uninformed plaintiff, makes a take-it-or-leave-it offer on the courthouse steps.[24] The informed defendant's settlement offer potentially signals his private information and the uninformed plaintiff must form Bayesian inferences when deciding how to respond to the offer.

Reinganum and Wilde (1986) characterize an elegant fully-separating equilibrium of this game where the defendant's offer perfectly reveals his type and the plaintiff mixes between accepting and rejecting the defendant's offer. The defendant's equilibrium settlement offer,

$$S(x) = \delta(x - c_p),$$

[24] Daughety and Reinganum (1993) present a model where both the timing of offers and the information structure are endogenous.

gives the plaintiff exactly the same payoff that she would get at trial.[25] The plaintiff subsequently randomizes between accepting and rejecting the offer where the probability that the plaintiff accepts $S(x)$ is given by

$$\pi(x) = e^{-(\bar{x}-x)/(c_p+c_d)}.[26]$$

Note that the probability of acceptance is increasing in the defendant's expected liability, x. This property is implied by the defendant's incentive compatibility constraint. Intuitively, the defendant must be rewarded in equilibrium for making higher settlement offers with a higher rate of acceptance by the plaintiff (who is indifferent between accepting and rejecting).

Mechanism design The screening and signaling models we have just seen are similar in some ways and very different in others. Importantly, the two models have similar implications for the selection of cases for trial. When the defendant has the private information then, on average, the cases that settle out of court have higher expected liability than the cases that go to trial. While the screening model generates this feature very starkly through the cutoff, $\hat{x}$, the signaling model generates it through the plaintiff's mixed strategy. (If the plaintiff had the private information instead, then this result would, of course, be reversed: cases that settle would have a lower expected liability than those that go to trial.) One difference is that, in the signaling model, every type of case (with the exception of type $\bar{x}$) has a positive probability of proceeding to trial in equilibrium.

An alternative approach to studying pretrial bargaining games is to consider the entire class of bargaining games using the Revelation Principle (Myerson, 1979). This principle tells us that any equilibrium of any extensive-form game may be represented as an outcome of a direct-revelation mechanism in which the defendant announces his type, $\tilde{x}$, and the mechanism designer subsequently maps this announcement into an outcome of the game. In the simplest settlement context, the mechanism would be a settlement offer, $S(\tilde{x})$, and a corresponding probability of settlement, $\pi(\tilde{x})$. Incentive compatibility for the defendant requires that he prefer to announce truthfully, and participation constraints imply that the players weakly prefer to play the direct revelation game than

[25] We will maintain the assumption that the plaintiff has a credible threat to take the case to trial regardless of his beliefs, i.e., $\underline{x} - c_p \geq 0$.

[26] Here is the derivation. If the defendant of type x were to mimic another type $\tilde{x}$ his expected payments would be $\pi(\tilde{x})\delta(\tilde{x} - c_p) + [1 - \pi(\tilde{x})]\delta(x + c_d)$. Incentive compatibility for a defendant of type x requires that he prefers to "tell the truth" and offer $S(x)$ rather than pretend to be someone who he is not by offering $S(\tilde{x})$. In other words, the derivative of this expression must be zero when $x = \tilde{x}$, or $\pi(x) - \pi'(x)(c_p + c_d) = 0$. The general solution of this differential equation is $\pi(x) = \alpha e^{x/(c_p+c_d)}$ where α is a positive constant. The boundary condition is $\pi(\bar{x}) = 1$. Suppose this boundary condition did not hold. The defendant of type $\bar{x}$ could raise his offer slightly above $\delta(\bar{x} - c_p)$ and the plaintiff would accept with certainty. (This offer dominates going to trial for any beliefs that the plaintiff might have about the defendant's type.) The D1 refinement of Cho and Kreps (1987) may be used to get rid of pooling equilibria.

go to trial. Both the screening and the signaling model discussed can be recast in this conceptual framework.

The mechanism-design approach provides us with proof that some cases will *necessarily* go to trial when the litigation costs are not too large. Formally, there does not exist a direct-revelation mechanism where all cases settle out of court. This may be easily established with a "proof by contradiction." Suppose that $\pi(\tilde{x}) = 1$ for all announced types—every case settles out of court. This implies that all defendant types must settle for exactly the same amount, call it S^*. If not, then the defendants, regardless of their information, would imitate the type with the lowest assigned settlement amount. The defendant's participation constraint requires that $S^* \leq x + c_d$ for all values of x, including the very lowest. (The defendant weakly prefers playing the mechanism to going to trial.) The participation constraint for the plaintiff requires that the plaintiff do better in settlement than she would do on average at trial, $S^* \geq E(x) - c_p$. Taken together, the value S^* can exist only if $E(x) - \underline{x} \leq c_p + c_d$, an assumption clearly violated when $c_p + c_d$ are small.

This result may be contrasted with that of Myerson and Satterthwaite's (1983) famous results for bilateral trade. They show in the bilateral-trade framework that if the buyer's valuation were private but the seller's type were known then a mechanism that set the price equal to the seller's cost would realize the gains from trade. In the litigation context, however, the plaintiff and defendant are unable to resolve their dispute even with one-sided incomplete information. It is also interesting to note that the breakdown occurs despite common knowledge that gains from trade exist: the litigants both know that they will jointly save $c_p + c_d$ by settling. In Myerson and Satterthwaite (1983), breakdowns only arise when the supports for the buyer's valuation and the seller's cost do not overlap, corresponding to the situation in which trade is not always efficient.

Although the literature has not delivered many positive or normative results about the entire class of pretrial bargaining games, several scholars have studied the mechanisms that achieve the Pareto frontier. For this so-called "optimal" mechanism, it can be shown that the selection effects described above hold here as well (so more liable defendants are more likely to settle). These mechanisms have also been investigated in more applied research. In particular, Spier (1994b) and Neeman and Klement (2005) use these mechanisms to consider the shifting of legal fees between winners and losers and pleadings rules, respectively. Although these mechanisms can provide a useful upper bound on private and social welfare, the approach has certain drawbacks. Unlike the bilateral trade mechanism of Myerson and Satterthwaite (1983) discussed above, the pretrial bargaining mechanisms are not necessarily implemented by standard extensive form bargaining games. In practice, these mechanisms would require players to commit to abide by the rules of the mechanism and for courts to enforce them.

2.2.3. Alternative frameworks

Mutual optimism Before the popularity and widespread adoption of techniques from information economics, law and economics scholars took a non-Bayesian approach to settlement breakdowns. Starting with the work of Landes (1971), Posner (1973), and Gould (1973), many scholars have taken the position that litigants may have different (and possibly inconsistent) priors about the outcome at trial. The plaintiff, for example, may believe the expected judgment at trial to be x_p while the defendant may believe it to be x_d. These divergent beliefs may arise when the two litigants receive different signals of the "true" expected damages, x, and may be influenced by their different backgrounds and experiences. Examining the bargaining zone, $[x_p - c_p, x_d + c_d]$, shows that the case will fail to settle when the plaintiff is much more optimistic than the defendant: $x_p - x_d > c_p + c_d$.

The optimism framework has two important advantages for applied work in litigation: its tractability and (arguably) its realism in many litigation settings. Scholars have used this framework to explore diverse topics such as the selection of cases for trial (Priest and Klein, 1984), fee-shifting (Shavell, 1982a), conflicts between lawyers and clients (Miller, 1987), and bifurcation of trials (Landes, 1993). It has also served as a foundation for empirical work on settlement (see Waldfogel, 1998). There is also interesting experimental and anecdotal evidence that litigants and their lawyers do tend to exhibit self-serving biases (Loewenstein et al., 1993). As a group, plaintiffs may have a tendency to overestimate the expected judgment at trial, $x_p > x$, while defendants as a group may tend to underestimate them, $x_p < x$. Indeed, these self-serving biases may serve as an advantage in bargaining—they allow the optimistic litigants to grab a greater share of the bargaining surplus—and can arise in evolutionary settings (Bar-Gill, 2006).

However, the optimism framework described above has disadvantages as well. In many ways, the optimistic litigants are "too stubborn": they stick with their inconsistent prior beliefs "come Hell or high water." In reality, many litigants—especially those with skilled lawyers—update their beliefs over time as new information emerges. They learn about the underlying merits of the case and are aware of strategies that their opponents employ. A careful understanding of this learning process is critical for both positive and normative analyses. It gives us a better understanding of the private litigation strategies we observe in reality and it is helpful in evaluating the effects of litigation reform. See Aumann (1976) for early work on common knowledge, Yildiz (2003, 2004) for recent theoretical work on learning and delay without common priors, and Watanabe (2005) for an empirical analysis.

Settlement externalities Cases may also fail to settle where there are externalities among existing claims.[27] To take a simple example, suppose that there are two plaintiffs and one defendant and that the payoff for each plaintiff at trial depends on how many cases ultimately go to trial. In this setting, settlement by one plaintiff changes the bargaining position and ultimate recovery of the second. Formally, suppose that each

[27] These types of externalities will be discussed in several topics in section 3, including joint and several liability and settlement with limited liability constraints.

plaintiff's expected payoff is x_2 if both plaintiffs go to trial and x_1 if only one plaintiff goes to trial. Depending upon the setting, x_2 may be either larger than or smaller than x_1.

First, suppose that $x_2 < x_1$. This may arise, for example, when the defendant enjoys economies of scale in case preparation. We can imagine settings where the defendant has a credible commitment to spend more on his defense when facing two plaintiffs than he has when fighting only one. Note that settlement by one plaintiff confers a positive externality on the second plaintiff, raising the payoff from x_2 to x_1. It is implausible that both plaintiffs will agree to settle out of court for x_2: if one plaintiff expects the other to settle out of court, he would rather reject the offer of x_2 and go to trial where he receives $x_1 > x_2$. Consequently, one or both plaintiffs will be able to command a settlement premium above and beyond x_2. It is not hard to see that, if the defendant's costs of litigation are not too large, then the defendant would rather forego settlement altogether than pay the plaintiffs a premium to compensate them for the positive externalities of settlement.[28]

Now suppose instead that $x_2 > x_1$. This could arise if the plaintiffs have a large joint fixed cost of pursuing their cases such as legal representation. Here, the plaintiffs are better off going to trial together rather than separately. Indeed, if one plaintiff were to settle then the second might decide to drop the case rather than pay a lawyer to pursue it independently. In contrast to the case where $x_1 > x_2$, here settlement by one plaintiff imposes a negative externality on the other. We would expect both cases to settle in this scenario. If the defendant offered each plaintiff $S = x_2$ (plus a penny) on the courthouse steps then both plaintiffs would be thrilled to accept. Indeed, the defendant may be able to induce them to settle for much less than that.[29]

Other reasons for bargaining failures There are many other reasons why settlement negotiations may fail. First, imagine that the object of litigation is indivisible—the custody of a child, for example—and that the litigants are liquidity-constrained. If the litigants value the object above and beyond the scope of monetary payment, then there is no scope for settlement. (See Shavell, 1993, and Mnookin and Kornhauser, 1979.) Second, one of the litigants (or their lawyer) may derive independent value from having his or her day in court (perhaps they enjoy the publicity!) or derive non-pecuniary pleasure from imposing costs on the opponent. Third, suppose that the parties have very

[28] The careful reader will notice that the defendant would try to "tie" the offers together: "My offer to settle for $S = x_2$ is good only if both of you accept. If only one accepts, the deal is off and I will take both of you to court." These types of offers are in fact observed in some class action settlements where a requirement is that a certain percentage of plaintiffs remain in the class.

[29] One way is to make settlement offers in sequence. There may be better strategies, however. If he offered between $S \in [x_1, x_2]$ then there are two equilibria: one where both plaintiffs accept and another where both reject. Although the latter is Pareto superior, the former is the risk-dominant outcome when S is not too small. See Spier (2002).

asymmetric stakes in the case. A corporate defendant, for example, may derive particular economic value from establishing a judgment in an early case that could chill the filing of future cases.

2.2.4. Normative implications

At first blush, there are strong normative arguments in favor of settlement. Take a single lawsuit—a personal injury case, perhaps—that would otherwise go to trial. As we have seen, it is certainly in the litigants' interest to resolve their dispute out of court. Through a private settlement, the parties can avoid their private litigation costs $c_p + c_d$ and (if they are risk averse) the risk premium associated with trials. It is also important to note that social costs are avoided as well—there are, of course, large fixed costs of maintaining the court system and legal infrastructure and significant marginal costs associated with any given trial (the judge's and jury's time, for example). *All else equal*, private settlement serves society's interest.

What makes this topic more interesting—and sometimes exceptionally challenging—is that *all else is not equal*. The rest of this section will discuss some of the real economic effects of settlement.

Primary incentives Suppose, for the sake of argument, that the potential defendant can take precautions at an *ex ante* stage to reduce the probability of an accident. Suppose further that all injured victims—the future plaintiffs—have positive expected value claims and that there is symmetric information during bargaining. These assumptions, while clearly unrealistic, allow us to isolate some of the basic effects of settlement on primary incentives.

Following an accident, the defendant is better off if he has the option to settle his claim. This is his revealed preference—if he were made worse off by settling he could simply refuse to settle and go to trial instead. Since the defendant anticipates settling on relatively advantageous terms, he has less incentive to take precautions to avoid the lawsuit to begin with. Simply put, settlement dilutes the defendant's incentives for care. See Polinsky and Rubinfeld (1988a). Perhaps surprisingly, the fact that the defendant takes less care is not necessarily a bad thing from a social welfare perspective. To see why, we will consider a specific example.

Suppose that damages are compensatory in the sense that x reflects the true harm that the plaintiff has suffered in the accident. If settlement were prohibited, then the socially optimal level of the defendant's precautions would reflect all of the private and public litigation costs associated with accidents in addition to the plaintiff's injuries. With compensatory damages, the defendant clearly underinvests relative to the social optimum—the defendant doesn't take into account all of the costs that the accident imposes on others, namely the plaintiff's litigation costs and society's costs of running the court system.

The desirability of settlement in this setting hinges on the defendant's bargaining power. Let's take the extreme situation where the plaintiff has all of the bargaining power: the defendant pays in settlement only slightly below what he would pay in total if the case went to trial, $x + c_d$. In anticipation of settlement, the defendant takes only slightly less care than before. In this case, settlement significantly reduces the litigation costs—a first-order improvement for social welfare—but incentives are not compromised. It follows that settlement enhances social welfare so long as the defendant does not have "too much" bargaining power. If, on the other hand, the defendant has significant power in negotiations then the effect on incentives could potentially outweigh the benefit of cost savings. The defendant, anticipating settling for $S = x - c_p$, would invest significantly below the first-best level. If $x - c_p$ is close to zero, for example, then the defendant has little incentive to take precautions at all.

Settlement can further dilute the defendant's incentives when asymmetric information is present. Recall that the screening model of settlement with asymmetric information featured some pooling of types. Consider, for example, a model where the uninformed plaintiff makes a final offer to a defendant who is privately informed about his liability for the accident. In equilibrium, defendants whose types (corresponding to expected liability) are above a threshold accept that offer. The pooling of defendant types may be bad for incentive reasons: the defendant has little marginal incentive to reduce his liability if he anticipates being in the pool. Spier (1997) presents a simple example along these lines where there is "too much" settlement in equilibrium. Incentives would be improved—and social welfare would be higher—if the plaintiff could commit to being tougher in settlement negotiations, making higher settlement offers to the defendant.[30]

Bringing suit The preceding discussion of primary incentives focused on the defendant's precaution decisions. We will now turn to the plaintiff's decision to bring suit. Following an accident, the plaintiff is made better off through settlement than she would be going to trial. Again, this is revealed preference—if the plaintiff were made worse off by settling she could simply refuse to settle and go to trial instead. Since the plaintiff expects to settle on relatively advantageous terms, she has a greater incentive to bring suit to begin with.[31] We will see that the plaintiff's increased incentive to pursue litigation may be either good or bad from a public policy perspective.

First, consider the effect of settlement on total litigation costs, taking the defendant's precautions as fixed. The ability to settle out of court surely reduces the private and social costs associated with *a given case*. (The cost will not be driven to zero, however; we have seen that asymmetric information is a robust obstacle to settlement and litigation costs are incurred in equilibrium.) At the same time, the ability to settle raises the *overall volume of cases* that are pursued. Indeed, the additional litigation costs associated

[30] In Spier's equilibrium, the defendant randomizes between taking due care and being negligent and the plaintiff randomizes between high and low settlement offers.

[31] This may be exacerbated when an uninformed defendant is making a final offer to a plaintiff who privately observes his damages. The screening equilibrium features a cutoff where plaintiffs with damages below the cutoff accept the offer. The pooling of plaintiff types may be bad for incentive reasons as well: plaintiffs who essentially have no injuries at all may succeed in extracting positive offers of settlement.

with the increase in cases may swamp the reduction in litigation costs associated with existing cases. In sum, settlement can lead the total litigation costs to either fall or rise.

Next, recall that the defendant's primary incentives are, of course, endogenous. Thus, a change in the plaintiff's incentives to file suit can generate additional indirect effects by affecting the defendant's incentives for care. Section 2.1.3 considered the social desirability of litigation, assuming that all cases that were filed went all the way to trial. We saw that there could be either too much litigation activity or too little litigation activity, depending on the nature of the feedback effects on the defendant's incentives. Bringing settlement into that analysis does not simplify matters. To illustrate, suppose that there were too many cases before allowing settlement because the costs of litigation were low relative to the benefit of improved incentives. Allowing settlement raises the number of cases further—a bad thing if we hold all else equal, even though the litigation costs associated with a given case fall.

Discussion The preceding discussion of the normative implications of settlement models showed that although there are a number of robust *positive implications* of settlement models, the *normative implications* depend very much on context. The well-established topics of nuisance suits, the shifting of legal fees, and accuracy—all topics that will be discussed in detail in section 3—raise normative issues along the lines discussed above. There is still more work to be done on the desirability of settlement from a *forward looking perspective*. Litigation is, by nature or design, a public good. Judicial decisions from early cases can influence the future in a variety of ways. In common law regimes, the opinions of judges—not the private settlement contracts—determine how the law itself evolves. Litigation can create social value by promoting the efficient evolution of laws to govern *future* economic activity. Trials may also create information that has independent economic value. Individuals who have suffered injuries may benefit directly from the groundwork laid by earlier claims.[32] The public also learns about the hazards of products and risky activities through the litigation activities of others.[33] A consumer, for example, may refrain from purchasing a risky product after observing the harms that similar products have caused to others.

3. Topics

3.1. Accuracy

It is generally thought that accuracy is valuable when imposing sanctions on offenders (whether they are criminals, tort offenders, or violators of private contracts). Intuitively,

[32] See also the discussion of secret settlements and the publicity effect in section 3.

[33] Hua and Spier (2005) consider a model where future actors use this information to fine-tune their accident avoidance behaviors. In general, the settling parties do not fully internalize the benefit that litigation has on these future actors.

if the legal rules are designed appropriately then the anticipation of accurate adjudication should create better incentives. Furthermore, accuracy should also help encourage innocent activities. Without it, individuals would tend to avoid engaging in value-creating activities that could be mistaken for violations and subsequently sanctioned. But accuracy does not come for free: there are significant private and public costs associated with designing more accurate legal rules. It is therefore important to identify circumstances where the additional costs of creating accurate outcomes are outweighed by the social benefits of accuracy.[34]

Suppose that an injurer (the potential defendant) must decide whether or not to engage in an activity. If the injurer takes no precautions, then the benefit to the injurer from engaging in the activity is b and the harm suffered by the victim (the potential plaintiff) is h. The injurer can avoid harming the victim if he invests e in precautions. The first-best outcome has the injurer taking precautions when his cost of taking them is smaller than the harm to the victim, $e < h$, and engaging in the activity when his private benefit is sufficiently high, $b \geq \min\{e, h\}$. This outcome could be obtained easily in a perfect world where litigation is accurate and free.[35] A damage rule that specifies compensatory damages where the victim is "made whole" would lead the injurer to make the correct cost-benefit tradeoff.

Suppose instead that liability is determined with error. While it is obvious whether or not the injurer engaged in the activity, the precaution taken by the injurer (and the associated harm borne by the victim) is not directly observable in a court of law. If the injurer took precautions, there is still a probability θ_1 that he will be held liable for damages, h. This is a "type 1 error"—the probability that an innocent person will be convicted. If the injurer failed to take precautions there is a probability θ_2 that he will get away with it. This is a "type 2 error."

Legal error distorts the injurer's decisions in this example. First, the injurer will take precautions if his expected cost from taking them, $e + \theta_1 h$, is smaller than his expected liability if he fails to take them, $(1 - \theta_2)h$. With a compensatory damage rule, the injurer will take precautions when $e < h(1 - \theta_1 - \theta_2)$. His decision is distorted because the type 1 error increases his cost of taking precautions while the type 2 error reduces his liability if he fails to take them. Second, the injurer will engage in the activity when $b \geq \min\{e + \theta_1 h, (1 - \theta_2)h\}$. Thus, the type 1 and type 2 errors clearly distort the level of economic activity as well.[36]

[34] See Kaplow and Shavell (1994) for an analysis of this tradeoff when liability is measured with error. See Kaplow (1998) for an excellent discussion of the many issues relating to accuracy more broadly.

[35] Polinsky and Shavell (1989) present a model where litigation is both inaccurate and costly. The litigation costs introduce an issue that is not addressed here: the decision of the victim to bring suit. They consider public policies such as imposing fines on losing plaintiffs to discourage their bringing suit.

[36] It is important to note that these distortions cannot be alleviated through a simple damage multiplier. A multiplier $\alpha = 1/[1 - \theta_1 - \theta_2]$ leads to the correct precaution incentives. However, this multiplier inflates the cost to the injurer of engaging in the risky activity to begin with, leading to a chill on economic activity. Additional instruments may include subsidies to those found innocent. See P'ng (1986).

To illustrate the problems associated with acquitting the guilty, suppose that $\theta_2 > 0$ and $\theta_1 = 0$ (so there is no chance that an innocent person will be convicted). This gives rise to two distortions: the injurer takes too few precautions and he engages in the activity too often. On the other hand, to illustrate the problems with convicting the innocent, suppose that $\theta_1 > 0$ and $\theta_2 = 0$ (so there is no chance that a guilty person will be acquitted). As before, the injurer takes too few precautions. In contrast, however, he will now engage in the activity too little rather than too much. The type 2 error—the chance that an innocent person will be convicted—discourages the relevant economic activity.

The previous example took the victim's harm level, h, as fixed. In reality, the injuries that arise due to the negligence of injurers are stochastic in nature. First, the injurer's failure to take precautions will not always cause an accident. Second, the victim's damages conditional upon an accident occurring are variable as well. Kaplow and Shavell (1996) argue that the *ex post* accurate verification of the victim's harm level may, or may not, be socially valuable in this setting. Accuracy is valuable if the injurer knew (or should have known) the victim's damages at the time when he chose his precaution level: the anticipation of an accurate award *ex post* gets the injurer to make the correct tradeoff *ex ante*. Accuracy is not valuable, however, if the victim's damages are purely stochastic and could not have been known by the injurer *ex ante*. There is no loss from setting the damage award equal to the expected or average harm in this case.[37] Indeed, when litigation is costly it is socially wasteful to devote resources to accurate outcomes when there is no corresponding benefit.

This section has highlighted two benefits of accuracy: deterrence and the encouragement of innocent economic endeavors. There are other benefits of accuracy as well that are important, but which fall outside this simple taxonomy. Accuracy may be valuable for risk-sharing reasons. If our victim above is risk averse, an accurate system that makes the victim whole is socially valuable because it reduces the risk premium that the victim would otherwise bear. Accuracy is also socially valuable when it creates better information, and therefore better incentives, for future actors. In a repeated litigation environment, the information created by earlier trials may help actors fine-tune their actions in the future.[38] Finally, the anticipation of precise investigations in the future will get injurers to do their homework ahead of time—they have a strong incentive to learn about the injuries that their risky activities cause.[39]

[37] Spier (1994c) argues that accuracy is valuable when precautions affect the magnitude as well as the probability of an accident.

[38] Hua and Spier (2005).

[39] Kaplow and Shavell (1992).

3.2. Evidence

3.2.1. The burden of proof

Black's Law Dictionary defines the burden of proof as a "legal device that operates in the absence of other proof to require that certain inferences be drawn from the available evidence ..." It is useful to think about this burden as having two parts. First, there is the so-called "burden of production" where a party must present sufficient evidence in favor or his claim or risk automatically losing without a full trial. Second, there is the burden of persuasion (or standard of proof) which provides the judge or jury with guidelines for how strong the evidence must be in order to find for the plaintiff (or prosecutor in a criminal proceeding). This burden is typically described in qualitative terms: "beyond a reasonable doubt," "clear and convincing evidence," "preponderance of the evidence," etc. A favorable decision for the plaintiff under the preponderance of the evidence standard, for example, is the point where the judge or jury believes that it is more likely that the plaintiff is in the right. In Bayesian terms, this corresponds to a posterior tipping point of 50%. "Beyond a reasonable doubt" is more subjective—what is "reasonable" is in the eye of the beholder.

The topic of evidence can be studied with the frameworks and techniques of information economics. Consider the following basic moral hazard example. An agent chooses a level of effort, either low effort (e_L) or high effort (e_H), which is not directly observed by the principal. The agent's choice generates a stochastic signal, s, which is distributed according to density function $f(s, e)$. The signal is informative in the sense that higher signals strengthen the decision maker's posterior belief that the effort taken was e_L rather than e_H.[40] Formally, this corresponds to the monotone likelihood ratio condition.[41] In our legal example, the signal s may be interpreted as the "evidence" suggesting the defendant's guilt. Correspondingly, the burden of persuasion (or standard of proof) corresponds to a cutoff, s^*, where evidence below the cutoff leads to acquittal and evidence above the cutoff leads to conviction.

This framework yields a stark implication for the design of both evidentiary standards (s^*) and the level of sanctions (D). Suppose that the cost of high effort is 1 and the cost of low effort is 0. The defendant will choose high effort so long as his cost of choosing high effort, 1, is smaller than the reduction in the expected sanction associated with the higher effort level,

$$1 \leq [F(s^*, e_L) - F(s^*, e_H)]D.$$

[40] Readers familiar with principal-agent models will note a subtle difference, here. In this example, a higher signal is associated with greater evidence of guilt, and hence *lower* effort on the part of the defendant. This is contrary to most standard models where higher signals are generally associated with *higher* effort. This distinction is highlighted to minimize confusion.

[41] Formally, the monotone likelihood ratio condition holds that $f(s, e_L)/f(s, e_H)$ is increasing in s for $e_L < e_H$.

Any pair of policy instruments, s^* and D, that satisfy this expression will provide the defendant with adequate incentives for care. But there are additional concerns associated with wrongful convictions. The monotone likelihood ratio property (MLRP) implies that by simultaneously raising the standard of proof, s^*, and the sanctions, D, perfect deterrence may be achieved at zero social cost. This is, of course, a variant of Becker's (1968) famous enforcement result—that penalties should be maximal while very little money should be spent apprehending criminals.

There are several reasons why we do not observe these extreme schemes in practice. First, civil sanctions are limited by a defendant's wealth and criminal sanctions are limited by a defendant's remaining lifetime. A significant increased probability of conviction is required to maintain incentives.[42] Second, juries may be unwilling to convict defendants when the sanctions are very high. Indeed, the subjective nature of criminal standards of proof (e.g., "beyond a reasonable doubt") gives the decision maker the discretion to define for himself what is reasonable. In Andreoni (1991), when sanctions are increased the jury convicts less often to avoid the higher cost of convicting the innocent (the cost of a type I error goes up since the innocent defendant is imprisoned longer). Since juries consequently convict less often, higher penalties may encourage rather than discourage crime. Third, the simple example above assumed that the evidence was exogenously generated. In practice, evidence is both costly to gather and subject to manipulation and misrepresentation.[43]

Rubinfeld and Sappington (1987) present a framework where the defendant chooses a level of litigation effort to influence the signal received by the court. They assume that defendants who are in fact innocent have "more productive" effort than their guilty counterparts. Innocent defendants consequently spend more money on their defense, endogenously leading to a better distribution of evidence at trial. The optimal social policy in this setting, which includes both the evidentiary standard (s^*) and the level of sanctions (D), will be chosen to balance the litigation costs and *ex ante* deterrence concerns. The defendant—whether innocent or guilty—will spend more money when the sanctions are higher and it is shown that the optimal policy has less than maximal sanctions.[44]

Sanchirico (1997) presents a model where plaintiffs, as well as defendants, make investments in their cases. Increasing the standard of proof for conviction has a good side benefit: plaintiffs who know that the defendant is innocent will choose to not file

[42] See Demougin and Fluet (2006) for an excellent discussion of optimal rules of evidence under wealth constraints. They argue that the penalty scheme that creates maximal incentives resembles the standard of preponderance of the evidence. Formally, rules that penalize an injurer when the evidence is "more likely" under negligence than due care (and conversely does not penalize the injurer when the evidence is "more likely" under due care) make the incentive compatibility constraint easier to satisfy.

[43] Other reasons include risk aversion, marginal deterrence, and heterogeneous injurers. See Bebchuk and Kaplow (1992) and the references therein.

[44] Their results are quite sensitive to the signaling technology of their model. See also related work by Sanchirico (2001).

suit, while plaintiffs who know that the defendant is guilty will litigate. In other words, the "self selection" may enhance social welfare. Bernardo et al. (2000) put additional structure on the litigation technology and find that, for evidentiary standards in an intermediate range, making a rule more pro-defendant can lead to more shirking and more litigation than before.

Hay and Spier (1997) assume that the body of evidence is fixed and not subject to manipulation. However, the two litigants may have different costs of acquiring or presenting the information. The burden of production may be viewed as assigning to one party the task of presenting the evidence to the court (and relieving his opponent to some extent of that task). Optimally used, this burden may minimize the expenditures devoted to gathering, presenting, and processing information in litigation. In practice, it is typical for plaintiffs to have the initial burden of producing enough evidence of the defendant's involvement in the case and indication of wrongdoing to justify proceeding. This makes sense in settings where there are innocent explanations for the plaintiff's injuries. Counterexamples exist, however. The taxpayer bears the burden of proof in income tax deficiency actions brought in the tax court by the IRS, for instance. This makes sense because the taxpayer would typically have greater access to evidence concerning his financial affairs than would the IRS.[45]

3.2.2. Disclosure and discovery

The basic framework presented in section 2 argued that asymmetric information between litigants about the likely outcome of the lawsuit could lead settlement talks to break down. The inefficiencies associated with bargaining impasses raise the question: "Why doesn't all of the information in the litigant's possession come out before trial?" Indeed, litigants will often voluntarily share information with each other before trial. An injured plaintiff, for example, may submit evidence of injury (x-rays, doctor's reports, etc.) to the defendant at the onset of filing to prove that she has a legitimate claim. In addition, in the United States and elsewhere there are laws that require litigants to disclose information when specifically requested to do so by the other side. An "interrogatory" is a set of questions that one side submits to the other side (and that must be answered) before trial. In a "deposition," lawyers may interview the other side's witnesses under oath.[46]

At an intuitive level, both voluntary and involuntary disclosure of evidence can serve important social objectives. First, the sharing of information before trial puts the parties on a more level playing field in the courtroom. This may well help to improve the accuracy of court decisions by making the game less one of rhetoric and more a fair contest based on the facts. Second, disclosure and discovery help to align the beliefs of

[45] See for example, Portillo v. Commissioner of Internal Revenue, 932 F2d 1128, 1133 (5th Cir. 1991).

[46] See Cooter and Rubinfeld (1994) for a good discussion of these institutions and a simple non-Bayesian model of disclosure where discovery can eliminate the "false optimism" of the litigants.

the parties about what will happen at trial, thereby facilitating private settlement. Critics of legal discovery point to its abuses, however, such as the ability of litigants to impose unfair costs on the other side. We will consider each of these issues in turn.

Trial outcomes Intuitively, both sides in a lawsuit have incentives to hide information that harms their case and to present evidence that helps their case. Discovery involves a set of formal rules and procedures that compel each side to share evidence that may, in the absence of discovery, never make it into the courtroom. It is thought that, in such a world, discovery may level the playing field by giving both sides access to the same information and thus improve the accuracy of legal decisions. While these ideas have some intuitive appeal, the formal models of disclosure of evidence can be subtle and often contradict this intuition.

Suppose, as in section 2, that a defendant privately observes his expected liability, represented by a parameter x drawn from a probability density function $f(x)$ on $[\underline{x}, \bar{x}]$. The defendant can costlessly and credibly reveal this information at trial. The court, a Bayesian player, does not receive any independent signals regarding the realization of x, but does know the distribution from which x is drawn. The defendant with the strongest case, $x = \underline{x}$, has an obvious incentive to disclose his innocence to the court and thereby secure an advantageous judgment. Indeed, since a Bayesian judge or jury would make an adverse inference if the defendant remained silent, the defendant with a slightly worse case, $x = \underline{x} + \Delta$, has an incentive to disclose this as well. This reasoning suggests that all information comes out at trial and an "accurate" outcome is obtained. This tendency for unraveling is familiar from the classic economic analyses of product quality and warranties (Grossman, 1981).

Grossman's unraveling result—that accuracy at trial is not compromised by the private incentives to withhold evidence—is sensitive to the underlying assumptions of the model.[47] First, not all defendants are able to reveal their private information—in many cases, hard evidence of innocence does not exist. Those defendants who do have hard evidence of guilt (they are in possession of a "smoking gun," perhaps) have an incentive to pool with the general population of defendant types who are simply unable to signal their cases' quality. See Shavell (1989a). Similarly, if the defendant has costs of disclosing evidence at trial then the unraveling will be incomplete in a world of voluntary disclosure.

With legal discovery, which mandates that the defendant must submit to interviews and answer questions before trial, it is more likely that the plaintiff will gain access to the information that the defendant would otherwise withhold. For example, the plaintiff's discovery activities may succeed in finding the "smoking gun," leading to the conviction of a defendant who otherwise would get off the hook. See Hay (1994) for further discussions and examples. In a complementary piece, Cooter and Rubinfeld (1994)

[47] See also Shin (1998), Shin (1994), and Sobel (1985).

argue that discovery may also improve accuracy as a consequence of "eliminating surprises" in the courtroom. Surprises at trial would typically lead to more spontaneous and less thought-out courtroom activities.[48]

Lewis and Poitevin (1997) present a model of costly disclosure to regulatory tribunals where disclosure creates a signal that is imperfectly correlated with the true state of affairs. With voluntary disclosure, they show that a litigant's decision to disclose information is itself a signal of strength: in equilibrium, strong litigants disclose and weak litigants do not (accepting the fact that non-disclosure will identify them as weak). Paradoxically, mandatory disclosure can reduce the accuracy of the ultimate court decision. When disclosure is voluntary, the weak types identify themselves by not sinking the costs of disclosure. When disclosure is mandatory, the court loses an important signal and some of the weak defendants may be exonerated.[49]

Settlement behavior To see the roles that disclosure and discovery play in private settlement negotiations, we return to the simple example from the section 2: A defendant privately observes his expected liability at trial, represented by a parameter x drawn from a probability density function $f(x)$ on $[\underline{x}, \bar{x}]$. Recall that if the uninformed plaintiff can make a single take-it-or-leave-it offer before trial (Bebchuk, 1984), the equilibrium will be characterized by a cutoff $\hat{x}$ and a settlement offer $S = \hat{x} + c_d$ where defendants whose types are above $\hat{x}$ will accept the offer to settle and those below will reject the offer and go to trial.

The game changes considerably when the defendant can credibly and voluntarily disclose his private information before the plaintiff makes the settlement offer. At first blush, one might think that the typical "unraveling" results would hold. A defendant who expects to pay the lowest damages at trial (type $\underline{x}$) is happy to disclose the information to the plaintiff and secures a low offer of settlement ($S = \underline{x} + c_d$). Continuing with this logic, it appears that defendants with slightly weaker cases will disclose as well. Perhaps surprisingly, complete unraveling does not arise in this setting. Shavell (1989b) shows that there is an equilibrium where defendants with types below $\hat{x}$ reveal their private information while those with types above $\hat{x}$ keep it hidden. The plaintiff, believing that the silent defendants come from the truncated distribution above $\hat{x}$ offers to settle with these types for $S = \hat{x} + c_d$. By remaining silent, a defendant with type

[48] See Cooter and Rubinfeld (1994, p. 446).

[49] Jost (1995) assumes that the defendant's disclosure is not credible in itself, but may be verified through discovery by the other side (which is analogous to costly auditing). He also takes the penalty structure, including penalties for misrepresentation, as exogenous. He shows that the defendant does not truthfully reveal his information in equilibrium. If he did, then the plaintiff would take everything he says at face value and not bother to spend time and money in discovery. And if the plaintiff does not audit, then the defendant would surely lie and pretend to have a strong case. Jost argues that it is better for a central authority, which has commitment power and a vested interest in long-run deterrence, to check the validity of the defendant's claims.

$x > \hat{x}$ earns information rents equal to $x - \hat{x}$.[50] Although unraveling is incomplete here, all cases ultimately settle out of court.

In contrast to the case of voluntary disclosure, all of the defendant's information would come to light if the plaintiff could somehow force the defendant to reveal his private information (through formal discovery channels, perhaps). By forcing the silent defendants to disclose their types, the plaintiff can tailor the settlement offer appropriately, offering $S = x + c_d$ instead of $S = \hat{x} + c_d$ to a defendant of type $x > \hat{x}$. To put it another way, the plaintiff can use discovery to "grab" the defendant's information rents of $x - \hat{x}$.

Shavell (1989b) shows that discovery plays a more important role when a fraction of the defendants are simply unable to disclose their private information before trial. As before, defendants who can prove that they have strong cases will voluntarily do so and receive low settlement offers. There is a group of defendants who remain silent, however: Some have strong cases but cannot credibly prove it to the plaintiff. Others have weak cases and decide to remain silent for strategic reasons. Mandatory disclosure plays an important role here because it "weeds out" the latter types—the weak defendants with credible information—from those defendants who are simply unable to reveal their information. Following discovery, the "silent" defendants have stronger cases on average, and so the plaintiff makes an even better offer than before. In this way, mandatory disclosure increases the rate of settlement.

Mnookin and Wilson (1998) present a model where discovery is both imperfect and expensive: each side can sink costs to get a more accurate signal of the opponent's type. These discovery efforts increase the probability of settlement by reducing the degree of asymmetric information. In their model, discovery is a public good: both the plaintiff and the defendant benefit (in expectation) from the discovery activities. The party engaging in the discovery benefits more, however, since one effect of discovery is to reduce the information rents captured by the other side.

There has been some empirical support for the idea that discovery facilitates settlement. Farber and White (1991) present an empirical analysis of 252 medical malpractice cases. Upon filing, it is likely that the plaintiffs were not well informed about the likelihood that the hospital or physician was negligent. Within this sample, 37% were dropped before trial, 58% were settled before trial, and 5% went to trial. Farber and White argue that the fact that so many cases were dropped following discovery indicates that the plaintiffs learned, through formal discovery channels, that the defendants

[50] The unraveling would be complete if the litigation costs were zero. See Hay (1994). If the defendant, rather than the plaintiff, made the settlement offer (as in Reinganum and Wilde's (1986) signaling model), then (with appropriate refinements) there could be complete unraveling even with positive litigation costs. The defendant with the strongest case would reveal it and offer $S = \underline{x} - c_p$ which would be accepted with probability one. In the absence of disclosure, this same offer would be accepted with a probability smaller than one. Sobel (1989) presents a model with two sided incomplete information where each litigant may be one of two types. He compares the set of equilibria that arise with mandatory discovery and no disclosure and discusses the voluntary disclosure case informally.

were not negligent. Similarly, the settlement of most of the remaining cases is consistent with the greater alignment of information following discovery.

The costs of discovery A private discovery request is, of course, costly and the costs of discovery are not typically internalized by the party who requests the information. Indeed, one side can force the other side to spend many months screening documents, sorting through private materials, and submitting to depositions. The potential for abuse here is obvious: discovery may be used as a strategic weapon. While the requesting party surely does not internalize all of the costs of discovery, he also does not internalize all of the benefits. Discovery, insofar as it increases the accuracy at trial, has the potential to create social benefits. The social benefits, which include increased deterrence and incentives for care, are enjoyed by society more broadly. It is therefore difficult to draw general conclusions about the desirability of discovery.[51]

Shepherd (1999) studies the time that litigants spend seeking discovery using a survey of attorneys in 369 federal civil cases.[52] He shows that defendants increased their discovery efforts, "tit-for-tat," in response to heightened discovery requests by the plaintiff. Interestingly, this "counterpunch" strategy was not observed for plaintiffs, who did not increase their requests in response to increased pressure from defendants.[53]

3.2.3. Admissibility of settlement negotiations at trial

Should the litigants' private settlement activities, including the offers that they make, their discovery requests, and the extent of their litigation expenditures, be admissible as evidence at trial? At first glance, one might assume that the answer is yes—after all, pretrial settlement activities can reveal valuable information. For example, in the section 2 we saw that litigants who possess more valuable information are more likely to forego settlement and litigate instead. Their failure to settle is an informative signal about the stakes of the claim. If the court has imprecise information to begin with, it can learn more about the truth—and consequently rule more accurately—when it can fully observe the settlement activities of the parties. It is perhaps a puzzle, then, that Rule 408 of the Federal Rule of Evidence in the United States prohibits the use of this information at trial.

Daughety and Reinganum (1995) provide an interesting economic rationale for the inadmissibility of settlement offers at trial: admissibility increases the rate of litigation.

[51] Schrag (1999) argues that placing limits on discovery reduces the costs of litigation and stimulates earlier settlement. For example, the 1983 revisions of rules 30, 31, and 33 of the Federal Rules of Civil Procedure placed limits on the number of depositions and interrogatories that each side could request of the others. In his model, increased discovery efforts do not unearth evidence *per se*, but instead influence the expected judgment directly.

[52] The results are similar when he also includes the time that litigants spend responding to discovery requests.

[53] These results suggest that the defendant's reaction curve slopes upward, while the plaintiff's slopes down. Shepherd (1999) also shows that hourly-fee attorneys made more discovery requests than their contingent-fee counterparts.

They extend the signaling model of Reinganum and Wilde (1986), where an informed plaintiff makes a take-it-or-leave-it offer, to include a Bayesian court. The court observes the settlement offer, updates its beliefs about the true state of the world and awards damages accordingly. As in Reinganum and Wilde (1986), there is a separating equilibrium where the plaintiff's offer reflects exactly the defendant's expected payments at trial. Incentive compatibility requires that the defendant mix between accepting and rejecting the offer, and that the probability of acceptance is decreasing in the size of the offer. When compared with the case of inadmissible settlement offers, the plaintiff has an incentive to "exaggerate" her offer and pretend to be of a higher type—by doing so, she influences the court and secures a higher damage award. Consequently, the defendant must reject with a higher probability than when settlements are inadmissible.[54] This argument implies that the litigation costs are lower when settlement offers are inadmissible at trial.

3.3. Sequential litigation

3.3.1. Appeals

An important characteristic of most legal systems is the right of a litigant who is dissatisfied with a lower court's decision to seek reconsideration by a higher court. This is true with civil, criminal, and administrative procedures in the United States as well as legal systems in other countries.

Shavell (1995) considers a stylized model where appeals are an efficient means of correcting the errors made at the lower court level. It does this by harnessing the private information of the litigants themselves. It is assumed that the litigants know whether a lower court ruling was in error or not and may launch a costly appeal. Shavell assumes that an incorrect decision is more likely to be overturned by the higher court than a correct decision. Litigants will tend to self-select in this environment: a litigant is more likely to sink the cost of appealing an earlier ruling if the probability of reversal—and hence the expected return from an appeal—is higher. This tendency to self-select is not perfect, however. When the cost of appeal is too low, then a litigant will appeal whether or not the lower court had rendered a correct decision. (Conversely, when the cost of appeal is sufficiently high, then no cases will be appealed.) By choosing an appropriate subsidy or tax, however, a social planner can align the litigants' appeal decisions with those of society more broadly.

An interesting implication of Shavell's analysis is that increasing the accuracy of lower court decisions is not a perfect substitute for the appeals process. "Increasing trial court accuracy reduces the frequency with which the appeals process is needed

[54] Kim and Ryu (2000) consider a model where the uninformed defendant makes the final settlement offer. They show that if the plaintiff's acceptance/rejection decision is admissible then the litigation rate will rise as well. Intuitively, the plaintiff would have an additional incentive to reject the offer to "convince" the court that she has a high type.

but not its desirability when errors are made."[55] The benefit of using the appeals system hinges on its ability to harness the information of the litigants themselves. By getting the litigants to self-select, resources tend to be spent on cases where a mistake has already been made. Similarly, the appeals system is not a perfect substitute for random audits performed by the upper level courts.[56]

It is important to note that Shavell's (1995) upper-level court is not a Bayesian decision maker. Indeed, if the upper level court were fully rational it would realize that only "mistakes" are appealed, and would therefore rule in favor of the appellant. This would, of course, interfere with Shavell's self-selection equilibrium: if the upper level always found in favor of the appellant, then all losers in the lower court—correct and incorrect decisions alike—would find it in their interest to appeal.

Daughety and Reinganum (2000a) consider a Bayesian model of appeals where the upper court perceives the private decision to appeal as informative and tries to rule "correctly" given its posterior beliefs. Formally, the authors assume that the both the appeals court and the litigants themselves receive private signals that are correlated with the truth. Technically, their signals are "affiliated" random variables (Milgrom and Weber, 1982). In equilibrium, a losing party appeals if and only if his or her signal exceeds a threshold. The upper court subsequently finds for the appellant and overturns the lower court's decision if and only if their own private signal exceeds another threshold.[57]

3.3.2. Bifurcation

Legal systems often feature a sequence of decisions before a final judgment is reached. Appeals systems, mentioned previously, allow for the re-litigation of an issue if one of the litigants is dissatisfied with a lower court's decision. In many other settings, there are sequential decisions on different issues before the final judgment. In criminal procedures, guilt is typically established before hearings to determine the convicted defendant's sentence take place. In products liability settings, proof must first be offered that the defendant was indeed the manufacturer of the product that caused the plaintiff harm before issues of negligence and damages can be considered.

[55] Shavell (1995, p. 387) notes that, in addition to efficiently reducing lower-court error through self-selection, the appeals process may also improve accuracy by providing lower court judges with better incentives to make careful and well-reasoned decisions. The idea is that judges dislike being reversed on appeal, and therefore they will devote more effort to their decisions and show less favoritism than otherwise.

[56] See Spitzer and Talley (2000) for a formal model of judicial auditing. They argue that higher levels of auditing are warranted when the lower courts are: (1) less accurate and (2) more swayed by ideology than by the facts of the case.

[57] Daughety and Reinganum (1999b) extend this logic to a horizontal sequence of courts facing similar cases, where each court along the chain receives an affiliated signal. While no appeals court's decision is precedential in another circuit, judges may view previous decisions in other circuits as a source of persuasive influence. They show that a herding phenomenon can arise where the earlier courts place more weight on their own private signals, wile the later courts discount their private signals in favor of the earlier courts' rulings. Herding is also referred to as an "informational cascade." See the survey by Bikchandani, Hirschleifer, and Welch (1998).

Landes (1993) presents the first formal analysis of the incentives to file, settle, and spend in bifurcated versus unitary trials.[58] In a "bifurcated" trial, the court first establishes the defendant's negligence before the plaintiff's damages are considered. In a "unitary" trial, the court determines both issues at the same time. His is an optimism framework, where the defendant and the plaintiff have potentially different estimates of the probability that the plaintiff will prevail on liability and of the damage award conditional upon liability being established.

When the private costs of establishing liability and damages are exogenous and settlement is not possible, then, conditional upon the plaintiff filing suit, bifurcation leads to lower litigation costs than a unitary trial does. The reason is simple: once the defendant is absolved of liability then the case is over and no further costs are incurred. Landes points out, however, that as a consequence of these litigation cost savings the plaintiff will file more suits than she otherwise would. Even though the litigation costs per case will fall, the number of cases will rise so that the overall effect on litigation costs is ambiguous.[59]

At first glance, it appears that bifurcation would have an added advantage over unitary trials when the plaintiff and defendant can settle their claims. The argument would go something like this: suppose that the plaintiff and defendant have the same assessment of the plaintiff's damages although they disagree about the defendant's liability. In a bifurcated proceeding, the parties would surely settle on damages following a plaintiff victory on liability. Therefore the litigation costs to establish damages in a unitary trial would avoided. The fallacy in this reasoning, as Landes points out, is that the parties could settle their damage dispute before a unitary trial as well, transforming a unitary trial in which both liability and damages are determined into one where only liability is considered. Therefore we would not expect that sequential and unitary trials would differ significantly on the propensity of private parties to settle.

Chen, Chien, and Chu (1997) reconsider Landes' questions in a model with asymmetric information instead of mutual optimism. In their model, the defendant is privately informed about both the probability that he will be found liable and the damages (conditional upon a finding for liability). With a unitary trial, plaintiffs run into difficulties making low settlement offers to the defendant. Since the cost of proceeding to a unitary trial is large, the plaintiff must maintain credibility not to drop the case following the rejection of a low settlement offer (as in Nalebuff, 1987). With a sequential trial, however, it is easier for the plaintiff to maintain credibility. Since the plaintiff will have another opportunity to settle before the stage in which damages are determined, the plaintiff's cost of proceeding is lower than in the case of a unitary trial. Chen, Chien, and Chu show that the overall effect on the settlement rate is ambiguous.

Finally, interesting insights are obtained when the costs of litigation are assumed to be a choice variable for the two parties. Landes (1993) provides a nice discussion of these

[58] However, there is a related informal discussion in Schwartz (1967).

[59] White (2002), in her analysis of asbestos trials, shows bifurcation raises the plaintiffs' expected returns and increases the number of cases that are filed.

issues, and Daughety and Reinganum (2000b) analyze and explicitly model endogenous litigation costs. It is clear, conditional upon the plaintiff winning in the liability stage, that both litigants will spend more in the damages stage. Intuitively, in a unitary trial the stakes for damages are smaller because the damages are discounted to reflect that chance that the plaintiff will lose on liability. According to Landes (1993), bifurcation transforms what was a fixed cost in a unitary trial into a variable cost in a bifurcated trial. The overall effect on the expected costs spent establishing damages is ambiguous, however, since the higher costs are borne less often in the bifurcated trial.

Interestingly, with endogenous litigation expenditures, bifurcated trials tend to favor defendants over plaintiffs.[60] To see why, consider the incentives of the two litigants to spend money in the liability stage of the bifurcated trial. Looking forward, the defendant has larger stakes than the plaintiff since the defendant expects to pay $x + c_d$ in total if he loses on liability while the plaintiff expects to receive $x - c_p$. To put this somewhat differently, the anticipated litigation costs in the damages stage drive a wedge between the plaintiff's and defendant's stakes in the liability stage. Consequently, the defendant marginal return from spending an extra dollar in the liability stage is higher than the plaintiff's and therefore the outcomes will be biased in favor of the defendant.

3.3.3. *Collateral estoppel*

A set of related rules and doctrines say when and whether a decision in one case will bind on another case *when at least one party is involved in both litigations*.[61] Under the doctrine of *Res Judicata*, the same claim between the same two parties may not be "re-litigated" (although parties may be permitted to appeal the outcome if they believe the court was in error). In criminal law, double jeopardy holds that a defendant cannot be tried twice for the same offense. Furthermore, the appeal rights under double jeopardy are asymmetric: the defendant has the right to appeal a conviction, but the prosecution may not appeal an acquittal. Related rules may apply when a single defendant (say) is facing a sequence of similar cases. For example, there may be several victims in an accident caused by a negligent truck driver. A finding of negligence in the first victim's case may bind for the second victim's case as well.

To explore some of the economic issues surrounding these rules, consider the following stylized example. A defendant, D, is facing a sequence of two plaintiffs with damages x_1 and x_2. Suppose that the two plaintiffs were injured in a highway collision with a truck and are bringing independent suits against the trucking company. The issues in both cases involve: (1) whether the trucking company was negligent and (2) the level of damages suffered by each plaintiff. The probability that the defendant will be found negligent by the court is p. We will consider two types of rules. The first, "2-sided

[60] See Landes (1993, pp. 22 and 41) and Daughety and Reinganum (2000b).

[61] This differs from precedent (although many of the economic issues are similar). First, precedent may hold even when the sequence of lawsuits involves different litigants. Second, it is often at the discretion of the judge whether to follow prior precedent or not (and perhaps create new precedent).

collateral estoppel," says that a finding of negligence in the first case would preclude relitigating the negligence issue in the second case. On the flip side, a finding that the defendant was not negligent in the first case would preclude the second plaintiff from bringing a case at all.

2-sided collateral estoppel Under the 2-sided rule described above, if the first plaintiff prevails, then the issue in the second case is one of determining damages alone. If the first plaintiff loses, then the second case is necessarily dismissed or dropped. This form of collateral estoppel makes a great deal of sense when the court's decision in the first case is unbiased and accurate. Litigation is expensive, after all, and it is a waste of resources to revisit the negligence issue once it has already been decided.

Several authors, including Spurr (1991), Katz (1988) and Che and Yi (1993) have shown that this rule leads to distortions when the court's decision in the first case depends on the litigation expenditures of the private parties. The defendant, as the long-run player, has higher stakes in the first case than does the first plaintiff—in fact, if the two plaintiffs have equal damages, then the defendant's stakes are twice as high. Consequently, the defendant's marginal return from additional litigation spending in the first case is higher. The divergence between the stakes, and the unequal litigation spending that results, will tend to bias the trial outcomes towards the defendant. As a result, the defendant's incentives to take care to begin with may be compromised. Note that this logic also suggests that the defendant would be more likely to appeal an adverse decision, thus exacerbating the bias. If given the choice, the defendant would also choose to litigate against the weaker plaintiff first. A plaintiff with low damages will spend less than a plaintiff with high damages, allowing the defendant to cheaply establish a favorable early ruling.

This two-sided rule could, potentially, influence the defendant's incentive to settle the first case. If the first case settles, then the court will hold the defendant liable with probability p if the second case goes to trial. Consequently, the settlement range before the second trial is $\{px_2 - c_p, px_2 + c_d\}$ where c_p and c_d are the litigation costs of the plaintiff and defendant, respectively. If, on the other hand, the first case goes to trial, then one of two scenarios will hold. If the defendant loses the first case, then the second plaintiff is sure to win on liability and the settlement range is $\{x_2 - c_p, x_2 + c_d\}$. If the defendant wins in the first round, then the second plaintiff has no grounds to bring a suit. There is no reason to think, *a priori*, that the defendant would derive a strategic benefit by settling with the first plaintiff rather than bringing the first case to trial. Indeed, if the plaintiff and defendant have equal litigation costs and equal bargaining power, then the "expected" settlement with the second plaintiff is px_2, whether the first case settles or not.

This result is sensitive to the particular assumptions about the bargaining power and litigation costs. If the defendant has all of the bargaining power, then he would prefer to settle the first claim. With all of the bargaining power, he can enjoy a settlement amount at the very bottom of the settlement range. If the first case settles, the defendant will settle the second claim for $S_2 = px_2 - c_p$. If the first case goes to trial and the defendant loses, the defendant subsequently offers to settle for $S_2 = x_2 - c_p$. The defendant clearly prefers the former scenario because the expected value of the latter is $p(x_2 - c_p)$. The opposite conclusion holds when the plaintiff has all of the bargaining power. The plaintiff in this instance enjoys a settlement at the very top of the settlement range. If the first case settles, the second plaintiff demands $S_2 = px_2 + c_d$ and the defendant accepts this demand. If the first case goes to trial and the defendant loses, the second case settles for $S_2 = x_2 + c_d$. In expectation, the defendant pays $p(x_2 + c_d)$ in this litigation scenario, which is less than if he settled the first case.[62]

Che and Yi (1993) explore the interaction between settlement incentives and asymmetric information with collateral estoppel. In their model, the two plaintiffs have private information about their damages and so bargaining vis-à-vis the second plaintiff will be inefficient: the second case will go to trial with positive probability. The 2-sided collateral estoppel rule has both a negative and a positive impact on the defendant's incentive to settle the first case.

First, if the defendant is found negligent in the first case, then the fundamental stakes of the second case will rise from px_2 to x_2. The severity of the asymmetric information problem will rise as well, since there is "more for the plaintiff and the defendant to disagree about."[63] The defendant suffers from the certainty of negligence in two ways: (1) he bears his litigation costs more often since negotiations are more likely to fail and (2) he shares more of the surplus with the second plaintiff (the plaintiff gets information rents). On the other hand, if the defendant is found not negligent in the first case, then the stakes of the second case drop to zero—the second case will not be brought. In this instance, there are no litigation costs incurred and the second plaintiff receives nothing in the way of information rents.[64]

1-sided collateral estoppel Suppose instead that the rule was one-sided: a finding of negligence would apply in the second case but a finding of adequate care would not. This rule clearly serves to benefit the later plaintiffs as the case against the defendant is "ratcheted up" over time. The defendant still has a long run stake in avoiding a finding of negligence in the first case as before. But now he lacks the benefit from a finding of no negligence in the first case. The over-spending effect mentioned above is still present although in a mitigated form (see Spurr, 1991).

[62] The astute reader will notice the litigation costs are being held constant as the stakes increase in this example. In reality, larger cases have larger costs of litigation associated with them. If the litigation costs were exactly proportional to the stakes, then there would be no difference between settlement and litigation from the defendant's perspective.

[63] Suppose that the probability of a finding of negligence is $1/2$ and that the second plaintiff's damages are uniformly distributed on the interval [100, 200]. If the first case settles, the stakes of the second case are distributed uniformly on the interval [50, 100]. Since the range of disagreement is smaller than before the second case is more likely to settle.

[64] Che and Yi also characterize conditions under which the settlement rate in the first round will rise or fall. The assumptions underlying their analysis are quite strong, however.

A short note in the Harvard Law Review (Note, 1992) considers the settlement effects of one-sided rules. Under these rules, the defendant clearly has a strong incentive to settle the first case. If he settles the first case, the stakes of the second case are simply px_2. If he litigates the first case and wins, the stakes of the second case are unchanged. But if he litigates the first case and loses, the stakes jump to x_2. Even absent the typical surplus created by litigation costs, there is a strong incentive for settlement in the first round. The paper argues that if the early plaintiffs are forward-looking and have bargaining power they will be able to extort more from the defendant in the settlement negotiations. Consequently, this one-sided rule will benefit early as well as later plaintiffs.

3.3.4. Precedent

A feature of Anglo-American legal systems is that legal rules can be created and changed by judges over time. The presence of earlier rulings on particular issues provides judges with a reason for ruling in the same way when new cases arise with similar issues. This section will mainly focus on the roles and incentives of judges in making laws over time. Private litigants with long-run interests also influence the evolution of the common law, of course.[65]

At first blush, adherence to the precedent set by earlier cases can create value for two reasons. First, past decisions embody useful information for future decision making, and so precedent will tend to lead to more accurate court decisions. (This is especially true if judges lack the expertise or the time to make accurate decisions in isolation.) Second, economic actors value the predictability that accompanies strong precedents. Predictability in a legal system will facilitate the smooth operation of an economy because it reduces the scope for disagreement. Predictable laws should correspond to fewer disagreements over liabilities, rights, and obligations and therefore produce fewer legal disputes. Predictability can also create greater value *ex ante* since economic agents are more likely to engage in productive activities when property rights are well-defined and secure.

Landes and Posner (1976) interpret precedent as a productive capital stock, or productive input into future court decisions. As with other capital stocks, it is argued that the value of a body of precedent depreciates over time. In the words of Posner (1992), "... accident law that was developed to deal with collisions between horse-drawn wagons will be less valuable applied to automobile collisions." Depreciation may come from technological obsolescence (as with horse-drawn wagons) of economic activities over time.[66] The common law gives judges the flexibility to make new rules and laws

[65] The preceding subsection discussed related issues in the context of collateral estoppel. Hay (1993) presents an analysis of preclusion rules and the relitigation of claims.

[66] Landes and Posner (1976) look at citations to previous cases in a sample of 658 federal appeals court decisions. Although citations are surely not a direct measure of precedent, they arguably serve as a useful proxy for the influence of past cases. The authors found, among other things, that the capital stock of precedent depreciates slower when: (1) there is less statutory activity in the area (so the formal written laws are not

to respond to these changes. This flexibility raises many questions. For instance, does the common law evolve efficiently? Does private settlement of disputes prevent efficient evolution? What if judges are self-interested and lazy?

Cooter et al. (1979) present a relatively early formal model of legal evolution. Their model is one of a negligence regime where the courts learn about—and subsequently adjust—the standards of care for injurers and victims. In their framework, the courts can observe the social welfare function and make incremental adjustments to improve the state of affairs. The authors show that the incremental adjustment process will converge to social efficiency. Their analysis makes use of many explicit and implicit assumptions, some of which are quite strong. In particular, they assume that cases are constantly litigated and that the courts have the knowledge of the underlying economic model needed to make wise decisions. They also assume that the decision makers—the judges—are benevolent and have the interests of society in mind. Although they provide a good starting point, these assumptions are unlikely to hold all of the time in practice.

As discussed throughout this chapter, the vast majority of cases that are filed ultimately settle out of court. Others are resolved before the plaintiff files a case at all, either through settlement or through a decision by the plaintiff not to pursue the case at all. This latter case—where the plaintiff drops the suit or fails to file it—will typically occur when the costs of going to trial are significant and the expected return is small. In a nutshell, the cases that actually make it before a judge—and whose decisions could serve as precedent for the future—are a very select group of cases. Rubin (1977) argues that this sample selection would actually *work in favor of efficiency*. (See also the related arguments of Priest, 1977.) The reasoning behind this type of argument is that inefficient laws lead to dead-weight losses in future economic activity. Therefore, private parties have a greater incentive to bring these cases in order to change future laws. This is especially true of private parties with long-run interests in changing these laws.[67]

Landes and Posner (1976) discuss the potential problems associated with judges making socially inefficient decisions while in the pursuit of their own preferences and political agendas. They argue that such tendencies are kept in check by a simple mechanism: a judge who makes a socially inefficient (but privately desirable) decision is more likely to be overruled in the future. Being overruled can have important consequences for the long run because being overruled on one case may well undermine the weight given to the judge's other decisions, reducing his or her citations and influence.[68]

changing) and (2) when the prior rulings were from the Supreme Court (which presumably selects for cases with more general impact). In particular, the citations to Supreme Court cases were on average twice as old for other courts (20 years versus 10 years old).

[67] In a model with endogenous litigation expenditures, Katz (1988) also argues that parties with long-run interests will spend more money in pursuit of changing inefficient laws and mitigating the associated dead-weight losses.

[68] Miceli and Cosgel (1994) present a formal model where judges have preferences over two things: the outcomes of individual cases and their "reputations." They show that a judge may well stick with prior precedent

Rasmusen (1994) formalizes some of the interactions among a sequence of judges. Judges have personal preferences over laws, and want to establish precedents that will be followed by others in the future. He shows that there can be multiple equilibria in this dynamic framework. In one equilibrium, all judges pursue their own private preferences by overturning past precedents. This brings the judge private value in the short run but not in the long run. Other equilibria exist, however, where judges cooperate with each other over time through "trigger strategies." In these equilibria, judges follow past precedent closely because violations would lead to future breakdowns where their own precedents would be violated by others. Schwartz (1992) and Kornhauser (1992) consider the incentives and strategies of tribunals, or multiple decision makers, who are interacting with each other both in the short run (on a given case) and in the long run. Judges may well engage in strategic behaviors not unlike those observed in the political arena (congressmen trading votes and the like).

3.4. *Allocating the costs of litigation*

3.4.1. *Loser-pays rules*

Note that the expected judgment at trial, x (as defined in section 2) can be interpreted as the product of the probability that the defendant will be found liable and the plaintiff's damages. In this section, we can normalize the damages to 1, making then x simply the probability that the plaintiff will win. We previously assumed that each side paid for its own costs of litigation, a rule that applies to most litigation in the United States. In contrast, with the so-called English Rule the loser must pay for the winner's legal expenses. The plaintiff's payoff at trial in this instance may be written $x - (1 - x)(c_p + c_d)$. The first term is the plaintiff's expected judgment at trial, x, and the second term reflects the plaintiff's expected litigation costs. With probability $(1 - x)$ the plaintiff loses and is forced to pay the defendant's legal costs as well as her own. By analogy, the defendant's expected payments at trial may be written $x + x(c_p + c_d)$.

If there is complete information about all of the relevant variables, the bargaining range is simply $[\underline{S}^{ER}, \overline{S}^{ER}] = [x - (1 - x)(c_p + c_d), x + x(c_p + c_d)]$. Notice that the size of the settlement range is exactly as it was for the American Rule in section 2.2: $\overline{S}^{ER} - \underline{S}^{ER} = c_p + c_d$. We will see that, compared with the American Rule, the English Rule has different implications for economic decisions. First, the English Rule changes the filing decisions of plaintiffs. Second, it affects the level of litigation spending. Fi-

against his personal preferences if the threat of reversal is sufficiently strong. He may deviate from precedent, however, if he views the outcome as sufficiently better for the case at hand or expects the decision to be upheld in the future. Levy (2005) presents a model where judges have career concerns and go against precedent to signal their abilities. See also recent work on precedent by Gennaioli and Shleifer (2005), Bustos (2006), Hylton (2006), Hadfield (2006), and Fon and Parisi (2003, 2004).

nally, it can change the litigation rate when the litigants' optimism and/or asymmetric information are taken into account.[69]

Filing decisions The English Rule changes the plaintiff's expected payoff at trial in a systematic way: it dilutes the value of low-probability-of-prevailing cases and enhances the value of high-probability-of-prevailing cases (Shavell, 1982a; Katz, 1990). This of course influences the plaintiff's decision about whether to pursue litigation to begin with. Assuming that all cases that are filed ultimately go to trial, a case is filed under the American Rule when the expected judgment exceeds the costs of litigation:

$$x > \tilde{x}^{AR} = c_p.$$

Under the English Rule, on the other hand, a case is pursued if and only if $x - (1 - x)(c_p + c_d) > 0$ or

$$x > \tilde{x}^{ER} = \frac{c_p + c_d}{1 + c_p + c_d}.$$

If $\tilde{x}^{AR} < \tilde{x}^{ER}$ then fewer cases are filed under the English Rule than under the American Rule; if $\tilde{x}^{AR} > \tilde{x}^{ER}$ then more cases are filed under the English Rule.[70]

To see this in a somewhat different way, notice that the plaintiff's expected litigation costs under the English Rule, $(1 - x)(c_p + c_d)$, exceed the litigation costs under the American Rule, c_p, if and only if $x < c_d/(c_p + c_d)$. This has important implications. At one extreme, when the plaintiff's probability of winning is low, then the English Rule can turn a viable case into a NEV proposition. In other words, the English Rule discourages low-probability-of-prevailing plaintiffs. At the other extreme, when x is high, the English Rule makes the plaintiff's case even stronger. In other words, the English Rule encourages high-probability-of-prevailing plaintiffs.[71]

Litigation spending Suppose that the litigation expenditures, c_p and c_d, affect the plaintiff's probability of success, x. Holding the defendant's expenditure fixed, the English Rule leads to greater litigation spending by the plaintiff for two reasons. First, there is an additional marginal benefit from greater spending since the stakes have increased from 1 to $1 + c_p + c_d$. Second, the marginal costs associated with spending are lower since the costs are partially externalized due to the fact that, under the English Rule, your opponent may be forced to pay your costs (Braeutigam, Owen, and Panzar, 1984; Hause, 1989; Katz, 1987).

[69] This discussion will focus on the theoretic literature. See Hughes and Snyder (1998) for a survey of the empirical literature in this area.

[70] $\tilde{x}^{AR} > \tilde{x}^{ER}$ if and only if $c_d < c_p^2/(1 - c_p)$.

[71] Kaplow (1993) and Polinsky and Rubinfeld (1996) discuss the normative implications of the English Rule and its effect on filing decisions. Polinsky and Rubinfeld discuss a more general set of rules that impose a penalty on losing plaintiffs and give a reward to winning plaintiffs.

Settlement behavior Suppose that the defendant is privately informed about x and, in particular, that he privately observes the probability that he will be found liable and that the uninformed plaintiff could make a take-it-or-leave-it settlement offer to him. As described in the section 2, any given settlement offer, S, corresponds to a cutoff value $\hat{x}$ where $S = \delta[\hat{x} + \hat{x}(c_p + c_d)]$. Defendant types whose liability is above the cutoff accept the offer and those whose liability is below the cutoff reject the offer and go to trail. The best screening offer corresponds to the cutoff, x^{ER}, that maximizes the plaintiff's expected payoff:

$$\int_{\underline{x}}^{x^{ER}} \delta[x - (1-x)(c_p + c_d)]f(x)dx + \left[1 - F\left(x^{ER}\right)\right]\delta\left[x^{ER} + x^{ER}(c_p + c_d)\right].$$

This gives the first-order condition:

$$\frac{1 - F(x^{ER})}{f(x^{ER})} = \frac{c_p + c_d}{1 + c_p + c_d}.$$

Comparing this expression to the first-order condition for the American Rule gives us a clear result: $x^{ER} > \hat{x}$. The cutoff under the English Rule is higher than the cutoff under the American Rule. It follows that more cases go to trial in equilibrium when the English Rule is adopted (Bebchuk, 1984). The intuition behind this result is simple. Trials occur in equilibrium because of asymmetric information about the outcome at trial—in this case, the defendant has private information about the probability that he will be found liable. This asymmetric information is exaggerated under the English Rule—now the parties are asymmetrically informed about who will bear the costs of litigation as well as the expected judgment at trial.[72] (Under the American Rule, the allocation of legal costs was common knowledge.) Intuitively, the English Rule creates more scope for disagreement between the two parties and the litigation rate consequently rises.[73]

The litigation rate will also rise with the English Rule when the informed defendant makes a take-it-or-leave-it offer to the plaintiff instead. As in section 2, we can construct a fully separating equilibrium with the following characteristic: the defendant signals his "type" through his offer of settlement, $S^{ER}(x) = \delta[x - (1-x)(c_p + c_d)]$ and the plaintiff randomizes between accepting and rejecting the offer. The interested reader can verify that an offer is accepted with probability:

$$\pi^{ER}(x) = e^{-(\bar{x}-x)\frac{(1+c_p+c_d)}{c_p+c_d}}.$$

[72] This result is driven by the fact that the defendant has private information about the probability of being held liable. If this probability were common knowledge and the parties were asymmetrically informed about the damages only, then the English Rule and the American Rule would lead to the same first-order conditions. Reinganum and Wilde (1986).

[73] Shavell (1982a) analyzes a model based on optimism. If parties are optimistic to begin with—so that the defendant estimates a lower value of x than the plaintiff estimates—then the English rule will exacerbate the problem. In this instance, the litigants are mutually optimistic about cost allocation in addition to damage awards.

Comparing this expression to the one for the American Rule shows that, given a defendant of type x, the acceptance probability is lower under the English Rule.[74] Again, the English Rule heightens the scope for disagreement: the defendant and the plaintiff now disagree about the allocation of costs in addition to the expected judgment at trial.

Several remarks are in order. First, the result that the English Rule raises the litigation rate may be reversed once litigation spending is taken into account. As mentioned above, we would expect the litigants to spend more money under the English Rule, and this will tend to create a greater incentive to settle. Second, the result is robust to changes in the information structure. If the plaintiff rather than the defendant were privately informed about his probability of prevailing, the litigation rate would still be higher under the English Rule. The selection of cases for trial would be reversed, of course: strong plaintiffs will tend to reject offers of settlement and go to trial instead. Polinsky and Rubinfeld (1998) point out that, in this case, the additional cases that go to trial under the English Rule all have lower probabilities of prevailing than those that go to trial under the American Rule. Finally, papers discussing the normative implications of the English Rule in a world of settlement include Hylton (2002), who analyzes a model with strict liability, and Spier (1997) who considers a negligence rule.

3.4.2. Offer-of-judgment rules

Another interesting class of fee-shifting rules bases the allocation of costs on the settlement offers that the litigants make to each other prior to trial. The most notable offer-of-judgment rule is Rule 68 of the United States Rules of Civil Procedure. Under this rule, if a plaintiff rejects a settlement offer made by the defendant and later receives a judgment that is less favorable than the offer, then the plaintiff is forced to bear the defendant's post-offer costs. Although Rule 68 is one-sided in the sense that it only applies to offers made by the defendant, other rules, such as California Code of Civil Procedure Rule 998, allow for two-sided cost shifting. Note that these offer-of-judgment rules are similar to the English Rule in the sense that they penalize the "loser" where the loser is defined relative to the settlement offers. Like the English Rule, offer-of-judgment rules can impact the likelihood of settlement; they can also affect the accuracy of settlements.

The litigation rate The effects of offer-of-judgment rules on litigation rates are interesting and subtle.[75] Spier (1994b) considers offer-of-judgment rules in a model where the defendant has private information. If the level of damages is common knowledge but there is disagreement over the probability of winning, then offer-of-judgment rules are very similar in theory to the English Rule. Since the settlement offer will typically

[74] This is a simple extension of Reinganum and Wilde (1986).

[75] Miller (1986) considers an optimism model. He showed that Rule 68 tends to be pro-defendant (which is not surprising since it is a one-sided rule) and has an ambiguous effect on the settlement rate.

lie between zero and the plaintiff's damage, the "loser" will end up picking up the expenses of the "winner." Not surprisingly, offer-of-judgment rules tend to increase the rate of litigation in these cases.[76]

Remarkably, this result may be reversed when liability is acknowledged but there is private information about damages: offer-of-judgment-rules discipline aggressive settlement offers and tend to lower the rate of litigation. To see why this is true, let us suppose as before that the defendant is privately informed about the expected judgment at trial, x.[77] The judgment at trial is noisy, however: $J = x + \varepsilon$ where ε is a noise term drawn from distribution $g(\varepsilon)$ with both mean and median equal to zero. Suppose that the plaintiff may make a single take-it-or-leave-it offer, S^J before trial. This offer will also be binding on the future allocation of costs: the defendant must bear the plaintiff's costs as well as his own if $J > S^J$ and the plaintiff will bear the defendant's costs if $J < S^J$. The defendant accepts the offer of judgment if and only if it is lower than what he would expect to pay at trial, $S^J < x + [1 - G(S^J - x)](c_p + c_d)$. This condition defines an implicit cutoff, $x(S^J)$ where defendants above the cutoff accept the offer and those below reject the offer and go to trial.[78] It is not difficult to extend the earlier analysis to show that the plaintiff's optimal offer satisfies the following first-order condition (Spier, 1994b):

$$1 - F\big(x\big(S^J\big)\big) - (c_p + c_d) f\big(x\big(S^J\big)\big) - (c_p + c_d) \int_{\underline{x}}^{x^J} g\big(S^J - x\big) f(x) dx = 0.$$

The intuition is straightforward. When the plaintiff raises the offer slightly there are both costs and benefits. $1 - F(x(S^J))$ represents the benefit: defendants with damages above $x(S^J)$ pay more in settlement. The next term represents a cost: when the offer rises, more costly trials occur. The third term represents an additional cost that is special to the offer-of-judgment rule: when the settlement offer is raised, it is more likely that the plaintiff will be forced to bear the costs should the case go to trial. It is through this third effect that offer-of-judgment rules discipline the plaintiff from making outrageous offers.

Settlement accuracy Bebchuk and Chang (1999) observe that offer-of-judgment rules can lead to greater accuracy in settlement. Their model is one of complete information—both parties observe the expected judgment at trial, x, and neither knows the realization of the noise term, ε. In their model, unlike the model of Spier (1994b), there are two

[76] Farmer and Pecorino (2000) show that this result does not necessarily hold when the level of damages is random. In this case, the analogy between the offer-of-judgment rule and the English rule no longer applies.

[77] Spier (1994b) considers an extensive form game where the plaintiff is privately informed instead. The identities of the informed and uninformed parties are reversed here to maintain consistency within the chapter. Spier also considers a more general mechanism design problem with two-sided private information and characterizes the fee-shifting rule and the bargaining game that maximizes the settlement rate. The fee-shifting rule has the same "flavor" as an offer-of-judgment rule.

[78] Note also that $x'(S^J) = 1$.

stages of bargaining: in the first stage, one of the two parties may make an offer of judgment, S^J, which is officially registered with the court. In the second stage, they arrive at the Nash bargaining solution (or, equivalently, they flip a coin to see who can make a take-it-or-leave it offer.

As a benchmark, recall that in the absence of fee-shifting, the bargaining range in the second stage would be $[x - c_p, x + c_d]$. With equal bargaining power the parties would settle at the midpoint: $S^* = x + (c_d - c_p)/2$. By backwards induction, this is what they would settle for in the first stage as well. Note that the party with the larger litigation cost is disadvantaged in the bargaining outcome. It is in this sense that the settlement does not accurately reflect the expected judgment at trial.

The outcome changes in a dramatic way with a two-sided offer-of-judgment rule. We will proceed by backwards induction. In the second stage, the most the defendant is willing to pay in settlement is $\overline{S} = x + [1 - G(S^J - x)](c_p + c_d)$ and the least the plaintiff is willing to accept is $\underline{S} = x - G(S^J - x)(c_p + c_d)$. With equal bargaining power, the parties would settle for an amount

$$\sigma\big(S^J\big) = x + \left[\frac{1}{2} - G(S^J - x)\right](c_p + c_d).$$

Working backwards to stage 1, suppose that the defendant must make the "offer of judgment," S^J. The defendant's equilibrium offer is clearly the value that satisfies $S^J = \sigma(S^J)$.[79] The same result would be obtained if the plaintiff rather than the defendant could make the offer of judgment. Using the assumption that the median $g(\varepsilon)$ is zero we have

$$S^J = x.$$

The parties settle for the expected judgment at trial, regardless of who has the higher litigation costs. The offer-of-judgment rule in this instance levels the playing field and generates a settlement outcome that accurately reflects the expected judgment at trial.

3.5. Negative expected value (NEV) claims and "frivolous litigation"

Many scholars and policy makers have expressed concerns about the presence of nuisance suits—that is, "frivolous" cases that are brought by aggressive plaintiffs for the sole purpose of extracting settlement offers from defendants. Despite the broad concern about nuisance suits and the importance of better understanding their sources and their ultimate control, there is little consensus about how a frivolous case should be defined. Some lawsuits, such as those where the plaintiff has suffered no damages or has no legal entitlement to recovery, are certainly frivolous. But not all frivolous cases are associated with a zero expected judgment—juries and judges may make mistakes and may

[79] If $S^J < \sigma(S^J)$ then the plaintiff would clearly reject S^J in the first round and subsequently settle for $\sigma(S^J)$ in the second stage.

grant an award despite the facts and the law. Furthermore, cases with very low expected judgments may, in fact, be socially valuable.

The law and economics literature has, for the most part, side-stepped this definitional problem by focusing instead on negative expected value (NEV) claims. We say that a plaintiff has an NEV claim *if the plaintiff's perceived return at the time of filing from taking the case all the way to trial is negative*. This corresponds to situations where the expected judgment is small when compared with the costs of filing, discovery, and litigation. At first blush, it does not appear that a plaintiff could profitably bring a lawsuit where the expected judgment at trial, x, is smaller than her costs of litigation, c_p. The plaintiff would surely choose to drop the case before trial! However, there are several reasons why NEV claims may in fact have positive settlement value. Also, there are several policy instruments that can affect these NEV claims.

3.5.1. Settlement of NEV claims

Bebchuk (1988) argues that the presence of asymmetric information—and the associated uncertainty—may make the plaintiff's threat to litigate the claim credible. In this model, as in Bebchuk (1984), the plaintiff privately observes the expected judgment at trial, x, filing is costless, and the defendant could make a take-it-or-leave-it offer before a costly trial. NEV cases occur with probability $F(c_p)$: the probability that x is smaller than the trial costs c_p. If $F(c_p)$ is small, the defendant will offer to settle for a positive amount: $S^* > 0$. Here, a plaintiff with a NEV claim benefits from the presence of other plaintiffs with positive expected value (PEV) suits. If $F(c_p)$ is large (so that there are many NEV claims) then the defendant will rationally refrain from settling—he offers $S^* = 0$ and all PEV plaintiffs go to trial.

Katz (1990) assumes that the plaintiff's costs are divided between the filing stage (k) and the trial stage ($c_p - k$) and that the defendant makes his settlement offer *after observing whether the plaintiff filed suit*. If $k < S^*$ then Bebchuk's (1988) result still holds: plaintiffs with NEV claims will file suit in the expectation—correct in equilibrium—that they will receive a favorable offer in the future. The equilibrium changes, however, when $k > S^*$. If the plaintiffs anticipated S^* as before, then no plaintiff would file suit with the intention of settling and, since only plaintiffs with PEV claims would find it worthwhile to file suit, the defendant would rationally make an offer above S^*. With filing costs taken into account, the defendant's offer becomes $S^{**} = k$. Plaintiffs with NEV claims are indifferent between filing suit and not and, in equilibrium, mix between these two options.[80]

Several papers have modeled the settlement of NEV claims in symmetric information environments. As discussed in section 2.2.1, Bebchuk (1996) shows that the divisibility of litigation costs can make the plaintiff's threat to litigate credible. Intuitively, the

[80] Although he considers this continuous case, Katz (1990) focuses on a two type example where the plaintiff is either injured or uninjured. In this example, the defendant's settlement offer is not deterministic as in the text but is a mixed strategy as well.

bulk of the costs have already been sunk once the case reaches the courthouse steps. Although the case may have begun with negative expected value it can be transformed into a positive expected value case when the costs of litigation are spread out over time. Rosenberg and Shavell (1985) show that plaintiffs may profitably extract settlement offers even when it is common knowledge that the plaintiff claims are NEV, and hence will be dropped before trial. Their argument rests not on divisibility but on an important timing assumption: the defendant had to sink some defense costs or risk a default or summary judgment before trial. The plaintiff is able to "hold up" the defendant in this scenario because the defendant is willing to pay *up to his defense costs* to settle the case.

3.5.2. Policy instruments

The ability of plaintiffs to extract settlements is affected in subtle ways by the fee shifting rules discussed earlier. First, suppose that a plaintiff's damages are substantial but the probability of winning is so low that the value of the case is only slightly above zero (it is a marginal PEV case). The English Rule dilutes the value of this case: the plaintiff will almost surely lose at trial and be forced to compensate the defendant for his litigation costs. This confirms the intuition the English rule serves to discourage low-probability-of-prevailing cases. Now suppose instead that the probability of winning is quite high but the damages are so small that the value of the case is slightly negative (it is a marginal NEV case). The English Rule will enhance the value of this case because the plaintiff's litigation costs will be shifted to the defendant at trial. Taken together, we see that the English rule can generate either more marginal cases or fewer of them, depending on the context.[81]

Some commentators have argued that contingent fees encourage frivolous claims and should therefore be prohibited. These arguments seem to be based on the idea that a plaintiff's threat to litigate is higher with contingent fees because it is the lawyer, not the plaintiff, who bears the direct costs of trials. In practice, however, the lawyer typically has more information about the case than the plaintiff and effectively controls the settlement decision. Since the lawyer will receive 33%, say, of the award but bear 100% of the costs, he would have an even greater incentive to avoid trials (Miller, 1987). Similarly, the lawyer would have even less of an incentive to represent an NEV case to begin with (Dana and Spier, 1993).

3.6. Contingent fees

In the United States, it is very common for plaintiffs to compensate their attorneys with contingent fees where the attorney receives a percentage of any settlement or judgment

[81] Bebchuk and Chang (1996) present an analysis of Rule 11 of the Federal Rules of Civil Procedure. In their framework, fee shifting is imposed not for winning or losing *per se*, but rather for deviations from case-specific thresholds for victory. They argue that Rule 11 may perform better at controlling frivolous lawsuits than alternatives.

but receives nothing if the case is lost. The typical contract involves a fixed percentage, often 33%, although there is variation. Some contracts specify that the percentage will decline with the amount of the award or the settlement, while others specify that a smaller percentage is received if the case is settled rather than litigated. See Rubinfeld and Scotchmer (1998) for an excellent discussion. Although contingent fees are most common in personal injury and medical malpractice cases, they appear in other types of litigation as well. It is much rarer to see contingent fees for defense attorneys, although they are occasionally adopted for tax cases. See the discussion in Dana and Spier (1993).

Contingent fees are, however, subject to restrictions and are often the focus of policy debates. In many European countries their use is totally prohibited. Despite their prevalence in the United States, the ABA Model Rules of Professional Conduct prohibit the outright purchase of cases by attorneys. The rationales behind these laws and restrictions tends to fall into one of two categories: (1) a paternalistic notion that unscrupulous attorneys can use contingent fees to take advantage of naïve plaintiffs or (2) the notion that contingent fees will lead to too high a level of litigation.[82]

In contrast to the views of policy makers, economists have tended to view contingent fees as a rational—and privately economically efficient—response to a variety of factors. First, contingent fees provide one way that liquidity-constrained plaintiffs can finance their cases. Without them, many plaintiffs would simply be priced out of the market. Second, contingent fees may provide for better risk sharing between plaintiffs and attorneys. This may be especially true if attorneys are "diversified" in the sense that they are representing many statistically-independent claims. Third, contingent fees help to overcome problems associated with moral hazard and asymmetric information. Fourth, contingent fees may (under some circumstances) serve as a valuable strategic commitment in negotiations. These latter issues are subtle and warrant a more detailed discussion.

3.6.1. Moral hazard

A moral hazard or hidden action problem exists when it is prohibitively costly for a principal (in this instance, the plaintiff) to directly monitor the effort chosen by the agent (here, the attorney). This type of problem is common in litigation since the plaintiff does not have the ability or the expertise to directly observe: (1) how much time the lawyer is spending on the case and (2) how hard the lawyer is working during those hours. Absent concerns for reputation or long-run play, paying the lawyer by the hour or by the job would clearly lead to insufficient attorney effort. Basing the attorney's pay on

[82] Santore and Viard (2001) have an alternative rationale for these limits: limits on the outright sale of claims generates more "rent" for the plaintiffs' attorneys as a group. The idea is this: absent prohibitions, attorneys will compete Bertrand-style and purchase claims for a large up-front fee and then receive 100% of the winnings. This scheme creates the efficient incentives for attorneys to work hard on the cases in the future, but gives them zero profits in an *ex ante* sense. With limits on the up-front transfer, the contingent percentage will fall short of 100%. Importantly, the individual rationality constraint will fail to bind.

the outcome of litigation, on the other hand, gives the attorney an incentive to work harder than he would otherwise. (See Schwartz and Mitchell, 1970 and Danzon, 1983 for early discussion of related issues.)

If attorney effort were the only concern, then an obvious economic solution exists: attorneys should "buy" cases from plaintiffs, paying an upfront fee to the plaintiff in exchange for 100% ownership of the settlement or judgment. Even if this were legal (and it is not), there would be practical reasons why this financial arrangement would not be universal. First, it is sometimes important to provide incentives to plaintiffs as well. If the plaintiffs were paid a lump-sum at the onset, they would have little incentive to cooperate with the lawyer or to put forth effort to maintain credibility on the stand. In short, the attorney and the plaintiff may be viewed as a "team," and sharing the outcome may be the second-best solution to the moral-hazard-in-teams problem. Second, we will see that the presence of *asymmetric information* may make selling the case undesirable for other reasons.

3.6.2. Asymmetric information

Plaintiffs and their attorneys typically have access to different information that is relevant to the case. The plaintiff will have first-hand knowledge of the extent of his or her injuries and the extent of contributory negligence. The attorney, on the other hand, knows more about his or her abilities and expertise in handling the case and knows more about the law that is relevant to the case. Suppose that the plaintiff can pinpoint the probability that the case will win at trial—it is either "high" or "low." The problem that the attorney faces in negotiating the contingent fee contract is analogous to the famous "Market for Lemons" problem: it is efficient for the attorney to "buy" the plaintiff's case (perhaps for risk-sharing or moral-hazard concerns) but he doesn't want to overpay for a "lemon" (i.e., a low-probability-of-winning case). Absent regulations on contingent fees, Rubinfeld and Scotchmer (1993) show that attorneys offer a menu of contracts in equilibrium: one with a high contingent percentage (and a relatively high purchase price) and the other with a low contingent percentage (and a lower purchase price). The client who believes that his case has a high chance of winning self-selects into the latter contract. Intuitively, the client signals the high quality of his case by his willingness to accept greater ownership of the case outcome. Rubinfeld and Scotchmer also characterize the equilibrium menu of contracts when the attorney has private information: the attorney signals his high quality through his willingness to accept contingent payment.

Dana and Spier (1993) show that contingent fees are the privately optimal financial arrangement when the attorney has better information than the plaintiff about the merits of the case. Intuitively, if the attorney were paid by the hour he would have little incentive to reveal to the client if the case lacked merit, and would pursue even claims with negative expected value.[83] Through the contingent fee, the plaintiff can be assured that

[83] Dana and Spier's (1993) model features equilibrium wages above the lawyers' opportunity cost of time. These attorney rents may result from fixed costs associated with overhead, case management, or perhaps

the attorney will make a more appropriate decision, pursuing *only cases that are more likely to win at trial*. This is an important insight and may be contrasted with the popular wisdom: in Dana and Spier (1993) contingent fees may lead to less (rather than more) litigation than otherwise.[84]

3.6.3. Settlement externalities

Using the same notation as in section 2, if the case goes to trial, the defendant will lose $x + c_d$ and the plaintiff and his attorney will (jointly) gain $x - c_p$.[85] As before, let x represent the plaintiff's expected judgment at trial and let c_p and c_d be litigation costs for the two sides. Absent agency issues, the settlement range would simply be $[\underline{S}, \bar{S}] = [x - c_p, x + c_d]$. The lower bound of the settlement range will change, however, depending upon the contingent fee received by the plaintiff's attorney, the allocation of costs, and who retains control over the settlement decision.

Formally, let α represent the attorney's fractional share of the joint cost (c_p) and let θ be the attorney's share of the judgment or settlement. If the case goes to trial, the lawyer's payoff will be $\theta x - \alpha c_p$ while the plaintiff's payoff will be $(1-\theta)x - (1-\alpha)c_p$. If the case settles out of court for S, on the other hand, the attorney's payoff is θS and the plaintiff's payoff is $(1-\theta)S$. The settlement offer that makes the attorney indifferent between settling out of court and going to trial is

$$\underline{S}^A = x - \left(\frac{\alpha}{\theta}\right)c_p,$$

while the offer that makes the plaintiff indifferent is

$$\underline{S}^P = x - \left(\frac{1-\alpha}{1-\theta}\right)c_p.$$

Clearly the lower bound on the settlement range depends crucially upon α, θ, and who retains control over the settlement decision.

First, suppose the proportion of the winnings received by the attorney is exactly equal to his share of the litigation costs, $\alpha = \theta$, then there is no conflict of interest between the plaintiff and his attorney: $\underline{S}^A = \underline{S}^P = \underline{S} = x - c_p$. The attorney will, for example, act in their joint interest when making settlement demands and responding to settlement offers. (This so-called "no-conflict" system, proposed by Polinsky and Rubinfeld,

education. Hourly fees would align the interests of lawyers and clients if they could be set to exactly equal the attorney's opportunity cost of time (see Emons, 2000). In that case, the lawyer would be indifferent between pursuing a case and not and there is an equilibrium where the lawyer makes the right decision on behalf of the client. If hourly fees could be fine-tuned in this way, they would be preferred to contingent fees.

[84] See also Miceli (1994). These theoretical results are consistent with the empirical evidence in Helland and Tabarrok (2003), who show that contingent fees are associated with higher-quality cases and a faster case resolution, and Danzon and Lillard (1983), who show a higher drop rate with contingent fees.

[85] We will abstract from the agency problem between the defendant and the defense attorney and assume that they always behave in their joint interest.

2003, also overcomes other agency problems, such as the effort problem identified in the discussion of moral hazard.)

More realistic, perhaps, is the case where $\alpha > \theta$. This would occur if the lawyer bears a disproportionate amount of the trial costs—$\alpha = 1$, perhaps—but only receives a third of the winnings—$\theta = 1/3$. It is easy to see in this case that the attorney's and the plaintiff's preferences diverge: $\underline{S}^A = x - 3c_p < \underline{S}$ and $\underline{S}^P = x > \underline{S}$. Their joint bargaining position is compromised if the attorney retains control over the settlement decision—the attorney is sorely tempted to settle for *too little*.[86, 87] (See Miller, 1987.) Giving the plaintiff control, on the other hand, enhances the bargaining outcome. Since the costs are "externalized" on the attorney, the plaintiff can credibly threaten to go to trial unless he receives an offer of at least $S = x$. (See Bebchuk and Guzman, 1996.) Put somewhat differently, contingent fees can sometimes serve as a bargaining tool in settlement negotiations.[88]

Hay (1997) analyzes a modified fee structure where the contingent fee associated with a settlement, θ_s, could differ from the fee associated with a judgment at trial, θ_t. In practice, it is not uncommon for the attorney to receive a greater percentage at trial: $\theta_t > \theta_s$. This fact is consistent with the idea that contingent fees are designed to both mitigate agency problems and to extract surplus during negotiations. The interested reader can easily verify that the lower bound of the settlement range, assuming the attorney retains control, is:

$$\underline{S}^A = \left(\frac{\theta_t}{\theta_s}\right)x - \left(\frac{\alpha}{\theta_s}\right)c_p.$$

If the plaintiff retains control, on the other hand, the lower bound is:

$$\underline{S}^P = \left(\frac{1-\theta_t}{1-\theta_s}\right)x - \left(\frac{1-\alpha}{1-\theta_s}\right)c_p.$$

If the plaintiff retains control over settlement decisions and has a great deal of bargaining power, then, Hay shows, the plaintiff will create a contract with a high contingent fee for trial winnings but low contingent fee for settlements. To take an extreme illustrative example, $\theta_t = 1$ creates a very powerful incentive for the attorney to work hard at trial. This raises the expected judgment at trial, and "scares" the defendant into accepting a high settlement offer. The plaintiff can benefit from this by setting a lower contingent fee for settlement. In the extreme, $\theta_s = 0$ allows the plaintiff to keep all of the money for himself! Interestingly, the same pattern emerges (although in modified form) when the attorney has all of the power at trial. θ_s cannot be too small, however, for if it were, then the settlement range would disappear.

[86] This discussion is assuming, of course, that the contingent fee contract is not renegotiated and that side payments between the attorney and client during settlement are impossible. In a frictionless world, they would certainly accept an offer of $S = x - c_p$.

[87] Thomason (1991) shows that plaintiffs who represent themselves in litigation have both higher settlement rates and higher settlement amounts, consistent with the theory.

[88] See also Choi (2003). Rickman (1999) who extends the dynamic asymmetric information model of Spier (1992a) to include the decisions of contingent-fee attorneys.

3.7. Tribunals

3.7.1. Judges and juries

Condorcet jury theorem Suppose there is a single defendant who is either innocent or guilty, and N jurors, each of whom privately receives an informative (but imperfect) signal about the defendant's guilt. Furthermore, suppose that the jurors have the same preferences: they would all agree on conviction versus acquittal if they each had access to all of the signals. More than 200 years ago, Condorcet (1785) proved that majority voting implements the jury's preferred outcome in the limit as N approaches infinity. An important underlying assumption in this analysis is that the jurors vote "sincerely" or "non-strategically"—they behave as if their vote, and only their vote, matters for the final outcome. This assumption has been maintained in most of the literature that followed.

In recent years, scholars have extended the analysis of jury decision making to include strategic behavior. To see why jurors would vote strategically, consider the following example, based on Austen-Smith and Banks (1996). There are three jurors who share a common prior that a defendant is in all likelihood innocent and would—absent additional evidence—unanimously vote to acquit. However, the jurors would change their minds and prefer to convict the defendant if (and only if) all three private signals indicate guilt. Suppose the first two jurors vote sincerely (a la Condorcet), entering a vote for acquittal when the signal is "innocent" and a vote for conviction when the signal is "guilty." The third juror, being rational and strategic, will not vote sincerely. With a majority rule, this third juror's vote is pivotal only when the votes of the first two jurors conflict. Therefore the third juror will vote to acquit even when his signal indicates guilt.

Strategic behavior among jurors can lead to unintended outcomes. Take, for example, the requirement in criminal courts that a defendant can be convicted only when the jurors are unanimous in their opinions. This rule is commonly thought to serve to reduce the probability of a type I error: the erroneous conviction of an innocent man. Although this is certainly true with sincere (non-strategic) voting, Feddersen and Pesendorfer (1998) have shown that the unanimity requirement can actually lead to more type I errors than a majority rule. Furthermore, in contrast to the limiting result in the Condorcet jury theorem, the probability of a type I error may actually increase with N, the number of jurors. Larger juries can also create a free-rider problem where jurors shirk in their individual responsibilities to pay attention and process information, reducing the accuracy of the ultimate decision (Mukhopadhaya, 2003).

Empirical work Casual observers tend to argue that judges may be preferred to juries because (1) juries are not capable of evaluating complex cases and (2) juries tend to be pro-plaintiff, granting inappropriately large awards for such things as punitive damages and pain and suffering. These premises have not been completely borne out in practice. Kalven and Zeisel (1966) survey judges who presided over civil jury trials and showed that the judges would have come to exactly the same decision as the jury in 78% of the cases. Importantly, the divergence was unbiased: the probability that the judge would have chosen to find a defendant liable when the jury found him not liable was the same as the chance that the judge would have found the defendant liable when the jury declared him not so. In an empirical analysis of awards, win rates, and settlements, Helland and Tabarrok (2000) find some confirmation of jury bias: conditional upon a finding of liability in their overall sample, a judge's award was on average only 31% of a jury award. They did find, however, that the plaintiff win rate was lower with a jury trial, and that controlling for a variety of case characteristics and selection bias got rid of most (but not all) of the jury bias.[89]

3.7.2. Adversarial versus inquisitorial systems

Scholars have found it useful to distinguish between *adversarial legal systems*, such as the one found in the United States, and *inquisitorial legal systems*, such as those found in continental Europe, Japan, and most other non-English speaking countries. In adversarial systems, the two sides in the dispute (or their representative agents) gather and process information. The two sides are self-interested and selective in what they reveal to each other and to the court.[90] In an inquisitorial system, the gathering and processing of evidence is centralized and often presided over by a judge.[91] Ideally, the inquisitor would be unbiased and hardworking, striving to uncover the truth. This distinction is, of course, a caricature—most legal regimes have elements of both.[92] Nevertheless, this stark distinction is useful for identifying the relevant costs and benefits of the different systems.

At first blush, inquisitorial systems have some obvious advantages over adversarial systems. In an adversarial process, the two sides gather evidence—evidence that is both helpful and harmful to their own position—and then choose to present the helpful information and discard the harmful information. Inquisitorial systems, by virtue of being centralized, can avoid the duplication (and wasted) efforts inherent in the adversarial process. Inquisitorial systems also appear to lead to more accurate trial outcomes: inquisitors, being disinterested parties, have no incentive to hide relevant information or mislead the decision maker.[93] Closer inspection, however, reveals that these critiques

[89] See also Clermont and Eisenberg (1992).

[90] The fact that adversarial systems such as that in the United States condone the hiding of damning information may appear puzzling at first. Why not require the two sides to reveal all of the information that they find? The reason may be quite simple: a rule that requires both sides to reveal all information would require a very sophisticated system of enforcement. In particular, there would have to be sanctions for not revealing all relevant information. In practice, it would be very difficult or impossible for the court to accurately identify what evidence should have been revealed. The errors in enforcement would likely have a chilling effect on the primary economic activity. See the discussion in Shavell (1989a).

[91] According to Posner (2003, table 1.1) the ratio of lawyers to judges in the United States is 55 to 1 while in Germany it is 7 to 1.

[92] See the discussion in Parisi (2002).

[93] See the informal arguments of Tullock (1980).

may be overstated and that there may be some distinct advantages of the adversarial system.

Milgrom and Roberts (1986) present a persuasion game where the parties can strategically withhold evidence at trial. The two parties to the lawsuit have equal access to all of the relevant evidence and can costlessly and credibly disclose it at trial. They find that the full information decision arises in equilibrium, even when the court is uninformed and strategically unsophisticated. The intuition is straightforward: since any piece of evidence will favor one party or the other surely, one party will have the private incentive to reveal it. In equilibrium, no stone is left unturned. In sum, accuracy is not necessarily compromised by the private incentives to withhold evidence.[94]

This stark result may no longer hold when parties have asymmetric access to evidence or when evidence is costly to gather and disclose. Daughety and Reinganum (2000b) present a model where the two parties independently engage in a sequential search for evidence. Although the court's decision depends on the information presented at trial, the court is not modeled as a purely Bayesian player. Instead, Daughety and Reinganum consider decision rules that satisfy axiomatic principles. Equilibrium bias may arise for several reasons in this setting. First, the two parties may have asymmetric sampling costs or access to very different sets of information. Furthermore, the parties may have very different stakes in the future. If the case is ultimately appealed, for example, then the costs of staging an appeal drive a wedge between what the two parties will win, creating a pro-defendant bias in the earlier stages.[95]

Dewatripont and Tirole (1999) argue that adversarial systems are more efficient than inquisitorial systems in providing incentives for information gathering. Suppose that two parties, A and B, are engaged in a dispute. An impartial and socially responsible judge must rule in one of three possible ways: finding for party A, finding for party B, and an intermediate outcome. Many types of disputes may fit this rough characterization. Suppose that A has been injured in an accident with B and that the doctrine of comparative negligence applies. If B was negligent but A took appropriate care then the judge would place full liability on B. On the other hand, if party A was negligent as well, then the judge may split the liability between the two parties. Custody battles between two parents provide another example: the judge can either award sole custody to one parent (excluding the other), or can award joint custody where the parents share the child.

[94] Shin (1998) argues that the adversarial system may be even more accurate. In his model, the two parties each receive an independent signal of the evidence, while the inquisitor receives one signal. The adversarial system may have an advantage in that the court may be able to glean better information in the event that nothing about the state of nature is directly revealed at trial.

[95] Froeb and Kobayashi (1996) argue that the full information decision may be obtained in the presence of costly information gathering even if the decision maker is fundamentally biased in favor of one of the parties. Intuitively, the favored party has less of an incentive to invest in information gathering and "free rides" on the court's bias. Despite being given an head start, the favored party may well use the same stopping rule as before, leading to exactly the same decision as would be made by an unbiased court.

Dewatripont and Tirole compare two regimes. In the inquisitorial system, an impartial (but possibly lazy) "inquisitor" is responsible for gathering evidence for both sides. In the adversarial system, partial advocates for each side are responsible for gathering the evidence. There may be evidence supporting each side of the case, but this evidence must be gathered before trial. (In the comparative negligence scenario, the injured party may be able to find proof of his or her care level. In the child custody scenario, a parent may be able to find proof of his or her suitability as a guardian.) In particular, if the gatherer of the information spends K there is a probability "x" that favorable evidence will be found. If only evidence favorable to A comes to light, the judge will find for A. Similarly, if the only evidence offered is favorable to B, then the judge will find for B. If both types of evidence are presented, then the judge will implement the intermediate outcome.

A crucial assumption in Dewatripont and Tirole's paper is that incentive schemes can be conditioned only upon the final judgment at trial and not upon the inquisitor or advocate's efforts or the evidence, directly. Recall that the intermediate judgment can arise because either: (1) the inquisitor was lazy and did not put in any effort into gathering evidence or (2) the inquisitor put in the effort but was unsuccessful at finding the evidence. The inability of the reward scheme to distinguish between these two alternatives dampens the incentives of the inquisitor. The adversarial system, on the other hand, provides better incentives for the advocates to find evidence.

The advantages of the adversarial system are easily seen in the following example where it is assumed that the agents cannot manipulate or distort evidence once it is gathered. We will assume that the gatherer of information receives a reward, W, only if the court finds exclusively for one of the two parties. Assuming that the advocate for party B will put in effort K to find evidence, party A will put in effort when:

$$x(1-x)W \geq K.$$

That is, when the probability that Advocate A finds evidence, x, multiplied by the probability that Advocate B does not find evidence, $1-x$, multiplied by the reward is greater than the cost of gathering evidence. Indeed, when $W > K/x(1-x)$ it is a dominant strategy for both advocates to gather evidence and the first best outcome is obtained.

Suppose instead that the evidence is gathered by a single inquisitor. Two incentive compatibility constraints are relevant here. The first is that the inquisitor prefers to gather evidence for both A and B rather than gathering no evidence at all. By analogy to the advocate case above, this constraint may be written:

$$2x(1-x)W \geq 2K.$$

The second constraint is that the inquisitor must prefer to gather evidence for both A and B (a payoff of $2x(1-x)W-2K$) rather than for just one party (a payoff of $xW-K$):

$$x(1-2x)W \geq K.$$

This second constraint is harder to satisfy than the first. Here is the intuition. If the inquisitor were to succeed in gathering evidence for party A, then any further effort on behalf of party B would be counterproductive. With evidence supporting both parties, the judge would rule for the "intermediate outcome" and the inquisitor would lose his reward, W.

It is not difficult to see that when $x \geq 1/2$ there does not exist a wage, W, that satisfies incentive compatibility for the inquisitor. In this case, the adversarial system is more efficient because it generates more information at trial. When $x < 1/2$, on the other hand, then it becomes necessary to pay rents to the inquisitor to get him to gather information favorable to both sides. Here, the adversarial system is more efficient because it generates the same information as the inquisitorial system but at lower cost to society. Dewatripont and Tirole confirm these same intuitions when the parties can conceal information from the court. Interestingly, when $x \geq 1/2$, then allowing the inquisitor to hide information (in order to avoid an intermediate ruling) creates better incentives—albeit at the expense of generating biased outcomes at trial.

The empirical work on this topic is scant. As part of a series comparing common- and civil-law countries, La Porta et al. (1997, 1998) provide evidence that common-law countries are better at protecting the rights of investors than civil-law countries. For example, they find that common-law countries had vastly more IPOs (initial public offerings) than those in the civil law tradition. The number of IPOs per million people was 2.2 for common-law countries, compared with .02 for countries in the French civil law tradition and .12 in the German tradition. Similarly, common-law countries had a much higher proportion of outsider-held stock as a fraction of GNP and many more firms per capita than their civil-law counterparts.[96]

3.7.3. Alternative dispute resolution (ADR)

Many cases are resolved outside of the legal arena through alternative dispute resolution (ADR). ADR is a very general term and encompasses both formal and informal proceedings that help parties resolve their disputes *outside* of formal litigation. Unlike settlement, which is typically achieved by the litigants themselves (and their agents—the lawyers—who represent them), ADR proceedings typically make use of third parties, including arbitrators, who offer opinions and/or advice.[97] ADR proceedings reflect two primary goals: (1) to reduce the transactions costs of reaching an agreement (these costs could include both the legal costs as well as the non-pecuniary costs to the litigants themselves) and (2) to come to "better" outcomes. Mnookin (1998) presents an excellent brief survey of ADR in practice and in theory. Although they tend to share common goals, ADR proceedings take on a wide variety of forms, each of which raises its own economic issues.

[96] Concentrated ownership and insider ownership may be good substitutes for legal protection.

[97] Mediation, where a third party may speak confidentially with the two sides and help them to find creative solutions to their dispute, is another common form of ADR. Since there has been little formal economic analysis of mediation, the topic has been downplayed here. See Ayres and Nalebuff (1997) for an economic model of mediation.

Some ADR systems are specified and designed by the parties themselves at either an *ex ante* or an *ex post* stage. Labor contracts, such as agreements between unions and employers, and commercial contracts, such as those between the wholesalers, dealers, and brokers of rough diamonds, include dispute resolution mechanisms to be used in the future if disputes arise.[98] Common features of these contracts include the ability of the parties to choose the arbitrators (the "third party") themselves and the binding nature of the third party's decision. Even if the parties did not specify the mechanism in the contract *ex ante* (perhaps because of drafting costs), they may still decide to adopt ADR *ex post* (after the dispute has arisen). Absent externalities and other market imperfections, the economist's view of *ex ante* agreements is that they are in the interest of society more broadly. Efficiency is served by dispute resolution mechanisms that reduce the *ex post* costs of disputes and facilitate economic activity *ex ante*.[99] This is not necessarily true for mechanisms that are chosen *ex post*, since *ex post* agreements would not naturally reflect the *ex ante* efficiency concerns. See Shavell (1994).

There are also many ADR systems that are adopted by the government instead of by the private parties. Many state courts in the United States, for example, have *mandatory* pretrial arbitration for automobile injury and medical malpractice cases. As with the private systems mentioned above, these court-annexed systems often allow the private parties to have some discretion in choosing the arbitrators and are less formal than trials.[100] Unlike the private systems, however, the court-annexed systems are typically non-binding: either party is free to reject the panel's decision and proceed to trial instead. Interestingly, some systems include fee-shifting provisions: if the litigant who rejected the arbitrator's decision receives a less favorable outcome at trial, then he or she will have to bear the opponent's post-arbitration legal costs. (See, for example, Farber and White, 1991.)

The empirical work on ADR has produced interesting and mixed results. Farber and White's (1991) study of medical malpractice claims suggests that pre-trial arbitration is informative to the private parties and that many cases settle subsequent to the arbitrator's decision. Yoon (2004) presents a time series difference-in-differences analysis of medical malpractice claims in Nevada and finds that fewer cases go to trial after the adoption of court-annexed ADR. This encouraging finding is dampened by Yoon's additional finding that ADR neither reduces the litigation costs nor shortens the delay to agreement. The relative dearth of research—both theoretical and empirical—makes ADR a ripe topic for further investigation.

[98] See the excellent case study and informal analysis of the New York Diamond Dealers Club in Bernstein (1992). In addition to specifying their own set of ADR procedures, the diamond industry also "opts out" of New York's contract law by specifying their own rules and codes.

[99] Dixit (2003) presents a model where an arbitrator—and expert—observes additional verifiable signals that are not observed by a court. The contracting parties benefit from this because they are able to write more complete contracts *ex ante*.

[100] See Bernstein (1993). Wittman (2003) uses California automobile cases to support his theory that litigants will choose arbitrators whose decision will reflect—and predict—the future outcome at trial. He finds little difference in the outcomes of cases that are arbitrated versus those that go to trial.

Finally, there is a small literature considering "final-offer arbitration" (FOA) where the arbitrator is bound to choose between final offers that are submitted by each side. FOA is particularly common in employer-union and baseball contract disputes. Farber (1980) characterizes the equilibrium strategies employed by the two sides when they share common uncertainty about the arbitrator's preferred outcome, exogenously given by $z \sim F(z)$. Given a final decision, x, the arbitrator's loss function is quadratic and given by $-(x-z)^2$. (If unconstrained, Farber's arbitrator would simply choose to implement his preferred outcome, $x = z$.) The players face a tradeoff when making their final offers: by making a more aggressive bid, the probability of acceptance is reduced (a cost) but the payoff conditional on the offer being chosen is higher (a benefit). In equilibrium, the plaintiff makes a higher offer than the defendant, but the average of the two offers equals the expected value of z.

Gibbons (1988) extended Farber's insights to include equilibrium learning by the arbitrator. In his normal learning model, there are two noisy signals of the underlying state of the world, z: one signal, s_A, is observed by the arbitrator and the other signal, s_P, is commonly observed by the disputants. As in Farber (1980), there is a separating equilibrium where the parties' final offers are *centered* on their private signal. The arbitrator consequently "learns" the disputants' signal from their offers and, together with his own signal, updates his beliefs about z.[101] Although not explicitly discussed by Gibbons, this model has an interesting implication for the accuracy of arbitration. In particular, although final offer arbitration allows the arbitrator to perfectly learn the signal s_P, it leads to an inefficient *ex post* decision since the arbitrator is bound to choose one of the final offers. With "conventional arbitration," however, the arbitrator is unconstrained. Gibbons shows that, as in final offer arbitration, there is a fully separating equilibrium where the arbitrator perfectly learns the parties' signal, s_P, and subsequently chooses the most efficient outcome. This, of course, begs the question of why we observe final offer arbitration to begin with.

3.8. *Multiparty litigation*

3.8.1. *Class actions*

When an injurer has harmed a group of victims, these victims may, under some circumstances, join their claims for the purpose of litigation and/or settlement. Consolidation has some obvious merits: the plaintiffs, the defendant, and the court may benefit from scale economies associated with common proceedings and legal representation (see Miller, 1998). This will obviously affect the private litigation incentives: plaintiffs who would otherwise have not pursued their claims are now able to do so, increasing the volume of litigation. The nature of settlement negotiations will change as well. A plaintiff

[101] See Farmer and Pecorino (2003) for a model that integrates FOA with settlement incentives. Ashenfelter et al. (1992) present experimental results comparing different arbitration systems and give other empirical references. See also Ashenfelter and Bloom (1984).

who joins a class will sacrifice, to a greater or lesser extent, her individuality. This manifests itself in a variety of ways. First, the plaintiff typically loses direct control over the attorney (who is now an agent for multiple parties) creating agency problems. Second, the outcome of a plaintiff's case typically rides on aggregate class characteristics or the characteristics of a "representative plaintiff" rather than her individual characteristics. Each of these issues will be discussed in turn.

Settling for coupons Many lawsuits over the years have settled for coupons as opposed to cash. This most commonly happens in antitrust suits, such as the price-fixing case against the airlines. There, consumers who could prove that they had traveled in the collusive period received (typically, non-tradable) coupons that could be applied to future flights. A main concern with these settlements is that relatively few of the coupons are ultimately redeemed in many of these suits. Consequently, the plaintiff class typically values the coupons at less than their face value. The lawyers representing the class, on the other hand, will often receive compensation that is based on the face value of the coupons. The coupons settlements could reflect an agency problem between the plaintiffs and the lawyer who is representing them. Since the class-action plaintiffs are not present at the bargaining table, the deal struck in settlement will naturally be susceptible to corruption.

In addition to the agency problems identified above, coupon settlements suffer from two additional flaws. First, as emphasized in Borenstein (1996) in a model with imperfect competition, the coupon settlement will typically affect the future pricing behavior of the defendants. Borenstein's main point can be illustrated in an example with undifferentiated Bertrand-style price competition. In the absence of coupons, the competitive price would settle at marginal cost and the defendants would make zero profits. If all consumers receive an abundant supply of coupons, the competitive price would be the marginal cost of production *plus the face value of a coupon*. At this price, firms earn zero profits. Note that the consumers are no better off with coupons than without: the competitive pricing has completely neutralized the effect of the coupons. With heterogeneous consumers, where some receive non-transferable coupons and others do not, we still see that consumers as a class do not benefit. The price of the product would rise to reflect that the coupon will be redeemed. Here, consumers who own the coupons will be better off than before while consumers who did not receive coupons will be worse off.

Polinsky and Rubinfeld (in preparation) zero in on an important welfare distortion arising from coupon settlements. Consumers are heterogeneous in their model, with stochastic per period demand that is uncorrelated across periods. Through a coupon settlement, consumers receive coupons in proportion to their earlier purchases during the so-called "injury period." In this context, several situations can arise. First suppose a consumer who had low demand in the injury period (and hence a small number of coupons) ends up with high demand later. This consumer will use all of his or her coupons and will also purchase additional tickets at full price: the consumer's purchase level does not hinge on the coupon's value. Suppose instead that a consumer who had high demand in the injury period (and hence has many coupons) has low demand subsequently. For this consumer, the coupons will increase his purchase level since the marginal unit is now cheaper. A fundamental distortion can arise in this case: the quantity that this consumer purchases may be inefficiently high in the sense that the marginal cost of the unit is higher than the intrinsic value that the consumer derives from its use.[102]

Settling for taxes In 1997, "The Tobacco Resolution" settled the lawsuits brought against big tobacco companies by the states (the main plaintiffs) for an expected $13 Billion per year in future tax revenues on cigarettes. The $13 Billion in tax revenues would not just come out of Big Tobacco's pockets, however: basic economic theory tells us that the market price of cigarettes would rise to reflect the tax and the burden would be borne (at least in part) by consumers. In the extreme, if the tobacco industry were perfectly competitive, then the entire burden would be borne by consumers. Why would the states and the tobacco companies agree to this settlement? As argued by Bulow and Klemperer (1998), Tobacco Resolution served the interests of the parties controlling the litigation: the state coffers were enhanced, the lawyers received a contingent fee tied to the tax revenues, and the tobacco companies effectively received license to collude and raise their prices. Bulow and Klemperer argue that, at the end of the day, only about $1 Billion per year would actually be borne by the defendants in that lawsuit. Since the big tobacco companies are able to externalize liability on consumers themselves, the deterrent benefit of liability is compromised.

3.8.2. Private incentives to consolidate

Damage averaging Che (1996) assumes that plaintiffs who join a class enjoy economies of scale from consolidation: the per-plaintiff cost of pursuing litigation decreases in the number of plaintiffs who join the class. This may be due, at least in part, to the streamlined proceedings. Instead of scrutinizing the damages of each individual plaintiff, the court may make a judgment based on the group average or, equivalently, on the damages of a randomly chosen plaintiff representative. Instead of receiving an award that is fine-tuned to his individual characteristics, a plaintiff's award reflects the average damages of the entire class.

Absent settlement, it is clear that plaintiffs with weak cases are more likely to join a class. A weak plaintiff has a stronger incentive to forego an individual hearing in order to receive an average judgment instead. Che shows that this adverse selection problem is mitigated when plaintiffs are privately informed about their damages. In short, weak

[102] A related distortion is also present in Borenstein's (1996) model, although the author does not highlight it. Borenstein also has heterogeneous consumers: some with coupons and others without. The consumers with the coupons may be over-consuming the product. This should be more likely when: (1) the products are less differentiated (so prices are lower), (2) when the proportion of consumers with coupons is relatively small, and (3) when the face value of the coupons is relatively high.

plaintiffs have an incentive to remain independent, too, in an attempt to "signal" that they have strong cases and, in equilibrium, fewer weak plaintiffs join the class. Although the ability to join class actions is shown to reduce the overall litigation spending it does not necessarily create a Pareto improvement.

Information rent extraction The presence of asymmetric information between the defendant and the plaintiff class, and among the plaintiffs themselves, implies that class formation can play a valuable strategic role. Che (2002) argues that classes may form to increase the members' bargaining power via information aggregation. In his model, the author assumes that the plaintiffs jointly evaluate a settlement offer and accept only if each and every member can be made better off than by going to trial. The commitment to a joint decision changes the defendant's choice of settlement offer. Instead of thinking about the distribution of an individual plaintiff's type, the defendant instead considers the distribution of the total damages (the sum of the individual damages). The defendant may subsequently choose to make more generous offers to the class as a whole than when bargaining one-on-one with individuals.[103] In this model, the plaintiffs also have an incentive to exaggerate their types among themselves in order to capture a greater share of the class settlement. This incentive to exaggerate commits the class to be even tougher and induces the defendant to make more generous settlement offers.

3.8.3. Joint and several liability

Just as it is common for a single injurer to harm many victims through his actions, there are many cases where a single victim is harmed by the actions of many injurers. The issue of how to allocate responsibility among multiple injurers has been a challenge for policy makers and scholars alike. Common rules include non-joint liability, where each losing defendant is held responsible only for his own share of the damages, and joint and several liability, where a single losing defendant can be held responsible for the entire level of the plaintiff's damages. While some proponents of joint and several liability have argued that the rule is good for public policy, the economic effects—especially those on settlement outcomes—are quite subtle.[104]

Kornhauser and Revesz (1994a, 1994b) show that settlement effects hinge on several factors including: (1) the treatment of prior settlements when determining the liability of a non-settling defendant and (2) the degree of correlation between the defendant's cases. They analyze the so-called unconditional *pro tanto* setoff rule.[105] With this rule,

[103] Suppose, for example, that there are two plaintiffs and each plaintiff's type (the expected return at trial) is either \$0 or \$1 with equal probability. Taken together, the two plaintiffs have damages \$0 with probability $1/4$, \$1 with probability $1/2$, and \$2 with probability $1/4$. A defendant who would have otherwise offered \$0 to each plaintiff individually may now find it in his interest to offer \$1 to the group.

[104] The survey of Kornhauser and Revesz (1998) also reviews the papers that abstract from the litigation process. How does joint and several liability perform in terms of deterrence? Other issues are addressed include the ability of defendants to collect from each other after an adverse judgment and the role of insolvency.

[105] Klerman (1996) shows that the results are actually sensitive to the formulation of the rule and compares some sensible alternatives.

the liability of a non-settling defendant is reduced, dollar for dollar, by the value of the previous settlements. Suppose, for example, that the plaintiff's damages are \$80 and there are two defendants. If the first defendant settles for S, the second defendant's liability is capped at \$80 $-S$. Kornhauser and Revesz show that joint and several liability encourages settlement when the two cases are positively correlated but discourages settlement when the two cases are independent.

We can see this in a concrete example. Let the unconditional probability of prevailing against each defendant be 50%. Suppose that the two cases are perfectly correlated: the two defendants will either lose together or win together at trial. Ignoring litigation costs, if both cases go to trial at the same time then the expected payment of each defendant is \$20. (They are held liable half the time and split the \$80 between them.) Now let's think about a settlement game. Suppose each defendant is presented with an offer to settle for $S = \$20$. If the first defendant accepts the offer then the second defendant's liability has changed: under the *pro tanto* setoff rule, the second defendant's liability is capped at \$80 − \$20 = \$60, which implies an expected judgment of \$30. The settlement decision of the first defendant has imposed a *negative externality* on the second. The plaintiff may be able to take advantage of this externality and induce the two defendants to settle out of court for more than \$20 apiece. The negative externality implies an additional incentive for settlement.[106]

Now suppose that the two cases are independent. If both cases go to trial, each defendant faces an expected judgment of \$30.[107] Putting aside the costs of litigation, suppose that the two defendants are presented with offers to settle for $S = \$30$. If the first defendant accepts the offer, then the second defendant's liability is capped at \$80 − \$30 = \$50, which implies an expected judgment of \$25. The settlement decision of the first defendant has imposed a *positive externality* on the second defendant. Note that the second defendant would refuse to settle on these terms and would demand (and receive) a discount. Kornhauser and Revesz show that, for low levels of correlation, the plaintiff will either reduce the settlement offers (if litigation costs are high) or forego settlement altogether and take both defendants to trial (if litigation costs are low).

This interesting theory has found some empirical support. Chang and Sigman (2000) test the results on settlement date for disputes between the Environmental Protection Agency (EPA) and Superfund defendants, the generators and transporters of hazardous waste and the owners of waste sites. They find that joint and several liability tends to promote settlement, and that the results are stronger for cases that are more highly correlated.

[106] Spier (1994a) argues that this may have bad normative implications as it may lead the defendant to overinvest in precautions *ex ante* and may encourage frivolous claims.

[107] A defendant loses alone and pays \$80 with probability $1/4$ and shares responsibility and pays \$40 with probability $1/4$.

3.8.4. Most-favored-nation clauses

Many settlement contracts in litigation involving multiple plaintiffs (or multiple defendants) include "most-favored-nation" (MFN) clauses. They work in the following way: if an early settlement agreement includes an MFN clause and the defendant settles later with another plaintiff for more money, the early settlers receive these terms, too.[108] It is important to emphasize that *these clauses typically apply to settlement payments only and not to judgments at trial*. This feature raises legitimate policy concerns. In the words of one critic, "Because [defendants] are 'straight-jacketed' by the most-favored-nations agreements with certain prior settling [plaintiffs], the strong public policies favoring complete settlement are being frustrated."[109]

Spier (2003a) argues that MFN clauses economize on delay costs when a single defendant makes repeated offers to plaintiffs who have private information about their prospects at trial. Without MFNs, recall that settlement negotiations resulted in cases settling on the courthouse steps (Spier, 1992a). These "11th hour" plaintiffs reject the defendant's early offers because they anticipate, correctly, that the defendant's offers could only improve with time. Through an MFN clause, the defendant induces these late-settling plaintiffs to accept early settlement offers instead.[110] While early settlement is socially desirable, there can sometimes be undesirable side effects of MFNs. In particular, the defendant may choose a more aggressive settlement strategy where more cases end up going to trial than before.

MFNs may also be used as an effective bargaining tool when future plaintiffs have the power to make offers as well. Intuitively, an MFN commits the defendant to be tough in future negotiations, placing an upper bound on what a future plaintiff can extract in settlement. The MFN allows the defendant and the early plaintiffs to capture a greater share of the future bargaining surplus. When committing to the MFN *ex ante*, the defendant and the early plaintiffs do not fully internalize the *ex post* cost of breakdowns, since at least part of that cost will be borne by the future plaintiffs. See Spier (2003b). In a model where early plaintiffs are also privately informed and signal their types through their settlement offers, Daughety and Reinganum (2004) show that MFNs make early settlement negotiations more efficient.[111] Taken together, the welfare effects of most-favored-nation clauses are ambiguous.

[108] These clauses have been included in prominent class action settlements (e.g., the 1999 Vitamins Antitrust case) as well as in settlement agreements with individual litigants (e.g., the tobacco settlements with the states of Florida, Mississippi, and Texas).

[109] In re Chicken Antitrust Litigation, 560 F. Supp 943 (Ga. 1979).

[110] These ideas are related to Butz (1990), where best-price provisions mitigate the time inconsistency problem that a monopolist faces when selling a durable good, thereby reducing social welfare. In contrast, here the reduction in delay improves social welfare.

[111] In essence, the potential for a future MFN payment lowers the incentive of an early plaintiff to inflate his or her settlement demand. This, in turn, reduces the need for the defendant to reject demands so as to ensure the separation of types.

3.8.5. Secret settlement

It is not uncommon for private litigants to settle their lawsuits quietly, where neither the existence of the suit nor the terms of the settlement are observable to the public. Secrecy may be facilitated through court-ordered sealing of the records, "gag orders," or through the parties themselves in their private contracts. Not surprisingly, secret settlements have attracted attention both in the policy arena and in academia. What are the advantages and disadvantages of allowing litigants to settle in secret? Secret settlements chill the flow of information to the public—information about the existence of legal remedies and the safety of products. There are important externalities at play: the value of this information accrues to future litigants and is not internalized by the settling parties.

Daughety and Reinganum (1999a, 2002) identify two types of information externalities. First, open settlements allow future plaintiffs to learn about the defendant's involvement in a case. A cancer patient, for example, may not know whether the cancer was a natural occurrence or whether the condition was caused by (or exacerbated by) a local waste site. This causal link among cases creates a "publicity effect" whereby an open (as opposed to secret) settlement makes it more likely that another plaintiff will file suit in the future. Second, an open settlement may provide future plaintiffs with information about the expected value of their claim—the "learning effect." Daughety and Reinganum (1999a) focus on the former publicity effect by assuming that the defendant's culpability—assumed to be private information of the defendant—is weakly correlated across plaintiffs. The uninformed plaintiff uses secret settlements as a screen: defendants who know that there are other victims are more likely to settle secretly. The plaintiff is able to extract "hush money" from these defendants in exchange for reducing the flow of information, money that is in effect coming out of the other victims' pockets. In this way, early plaintiffs can enrich themselves at the expense of later plaintiffs.[112] Importantly, secrecy can compromise firms' behavior and product safety choices in a market setting. Daughety and Reinganum (2005) present a signaling model where the average safety of consumer products is higher when firms can commit *ex ante* to settle all future disputes in the open.

3.9. Additional topics

3.9.1. Plea bargaining

Most criminal cases in the United States are resolved before trial through a process known as plea bargaining. The prosecutor and the defendant typically negotiate a guilty plea in exchange for a lesser charge (which is associated with a lighter sentence for

[112] Daughety and Reinganum (2002) introduce the learning effect by assuming strong correlation of culpability across cases. This complicates matters because there is an offsetting effect: as in section 2, defendants who are more culpable are more likely to settle and this creates an adverse inference on the part of future plaintiffs.

the defendant). In many ways, the economic approach to plea bargaining is similar to that of civil settlement. Many themes—including the avoidance of direct litigation costs and risks, the role of private information, and the feedback effects on deterrence and social welfare more broadly—are common to both. They diverge, however, in their assumptions about the prosecutor's preferences. In civil cases, the damages paid by the defendant are typically received by the plaintiff. In criminal cases the prosecutor does not "receive" the prison sentence directly.

The papers in the plea bargaining literature vary in their assumptions regarding the prosecutor's payoff function. Landes (1971), in the first formal analysis of plea bargaining, assumes that the prosecutor maximizes the sum of expected sentences subject to a resource constraint. His approach is more aligned with the models of civil litigation and settlement than the literature that follows. Several subsequent papers, including Grossman and Katz (1983) and Reinganum (1988), assume that the prosecutor represents the interests of society more broadly with a payoff function that includes type I and type II errors in addition to litigation costs.

Grossman and Katz assume that the defendant privately observes his guilt, which is correlated with the probability that he will be convicted at trial. The uninformed prosecutor can make a single take-it-or-leave-it offer of a reduced sentence in exchange for a guilty plea. Although they abstract from the direct costs of litigation, both the defendant and prosecutor are assumed to be risk averse and so, all else equal, they prefer to settle before trial. As in Bebchuk's (1984) analysis of civil settlement, the plea bargain offered by the prosecutor screens among the different defendant types: the guilty defendant is more likely to accept the offer than the defendant who is convinced of his innocence.[113]

Reinganum (1988) extends Grossman and Katz to allow the prosecutor to have better information about the probability of conviction at trial, a variable that is correlated with the defendant's innocence or guilt.[114] The prosecutor's offer of a plea bargain serves as a signal to the defendant who would subsequently update his beliefs about trial. The partial pooling equilibrium has the following form: when the probability of conviction is below a cutoff, then the prosecutor drops the case (formally, the sentence offered is zero). When the probability is above the cutoff, then the offered sentence perfectly reveals the prosecutor's private information and is increasing in the probability of conviction. As in the model of Reingaum and Wilde (1986), the defendant mixes between accepting and rejecting the prosecutor's offer with high sentences being rejected more often.[115] An important implication of this model, one that distinguishes it from Grossman and Katz (1983), is that trials are more likely when the defendant is guilty.

[113] Notice that the prosecutor knows that defendants who reject the plea are more likely to be innocent; thus, he has a greater incentive to drop charges after the plea offer is rejected. This observation is analogous to Nalebuff (1987). Baker and Mezzetti (2001) extend the model to take this into account.

[114] Only the prosecutor's private information is payoff-relevant to the defendant at trial, although the defendant's private information is relevant for the social welfare function (which corresponds to the prosecutor's payoff).

[115] This is an implication of incentive compatibility for the prosecutor.

Should prosecutors be granted full discretion in the offers that they make to defendants before trial? Reinganum (1988) shows that the answer may be "no," even when the prosecutor has the interests of society in mind *ex post*.[116] Reinganum compares the equilibrium described above to a regime where the prosecutor's offer cannot be fine-tuned to the signal received—in other words, a forced "pooling" offer. The equilibrium that results is akin to Grossman and Katz (1983): the pooling offer screens among the defendants where the innocent go to trial and the guilty accept the offer. If there are many guilty defendants in the overall population, then limiting prosecutorial discretion is socially desirable—with discretion, these guilty defendants are associated with costly trials. If there are many innocent defendants, on the other hand, then discretion is preferred. The benefit of the equilibrium described above is that cases against innocent defendants are likely to be dropped.[117]

3.9.2. Case selection and the 50% hypothesis

The cases that end up in trial are the "tip of the iceberg"—the vast majority of filed cases are settled before trial and even more are never filed to begin with. The cases that end up in the courtroom result from bargaining failures—failures that can arise for many reasons: asymmetric information, divergent expectations, the long-run interests of the parties (such as the need to establish precedent in a civil rights case), or a variety of externalities. In sum, the cases that go to trial are an unusual sample of cases and are likely to differ—perhaps systematically—from the cases that never reach the courtroom.

In 1984, Priest and Klein presented a "divergent-expectations" model of litigation and settlement with a striking prediction: for cases that go to trial, the probability of the plaintiff winning tends towards 50%. In contrast, for cases that settle out of court, the probability that the plaintiff would win (had they gone to trial instead) could be systematically higher or lower than 50%.[118] This paper has received a great deal of attention in the law and economics literature for (presumably) several reasons. First, they illustrate that litigated cases are unrepresentative of the broader case population. (Bebchuk, 1984, and others, of course, share this feature.) Second, the "50% hypothesis" is consistent with many people's rough intuition: if the evidence in a case *clearly* favors one party over the other, then the case should settle; if it is *unclear* who should win, there is greater scope for disagreement. Finally, the 50% hypothesis readily lends itself to empirical testing.

Shavell (1996) argues that any plaintiff win rate at trial is possible under more general assumptions. Most obviously, if both parties firmly believe that the probability that the

[116] Reinganum (2000) presents a signaling model where an informed defendant makes an offer to an uninformed prosecutor. In this model, the prosecutor's incentives are not aligned with society's incentives.

[117] To put it somewhat differently, the prosecutor may be the victim of his own discretion. The signaling equilibrium that results from discretion is socially inefficient. The prosecutor is better off if his hands are tied.

[118] Waldfogel (1998) gives a clear presentation of Priest and Klein's assumptions and results and surveys the related empirical literature. Asymmetric stakes, including situations where one side has a stronger interest in the case than the other, would change the results.

plaintiff will prevail is bounded between, say, 0 and $1/3$, then the plaintiff win rates for both settled and litigated cases must be below $1/2$. Less obviously, one can construct examples with asymmetric information that generate: (1) any given win rate at trial, p^T, and (2) any given (implied) win rate for settled cases p^S. To see why this is true, suppose that the distribution of plaintiff win- rates, $f(p)$, has full support on [0, 1]. We know from the earlier presentation of the basic framework that if the plaintiff has private information about her winning probability then plaintiffs with strong cases (high plaintiff win rates) are more likely to go to trial than those with weak cases (low plaintiff win rates). This information structure is consistent with examples where $p^S < p^T$. One can construct a distribution $f(p)$, litigation costs, c_p and c_d, and an extensive form game consistent with this outcome. If, on the other hand, the defendant privately observes the probability winning, then the selection goes in the other direction: defendants with strong cases (low plaintiff win rates) are more likely to go to trial than those with weak cases (high plaintiff win rates). This information structure is consistent with $p^S > p^T$ and, as above, consistent examples can be constructed.

Priest and Klein's 50% hypothesis was a limiting result where the errors in observation become increasingly precise—in the limit, parties beliefs converge and trial rates approach zero. Given the strong assumptions needed to generate the 50% result, it is not surprising that there is little empirical result for the strong form of the hypothesis.[119] More generally, however, Priest and Klein's framework suggests that trial rates may be systematically related to plaintiff win rates. Empirical work along these lines has been more fruitful. Waldfogel (1995), for example, empirically documents the relationship between litigation rates and plaintiff win rates in a structural model using data from broad variety of cases (including tort, contract, property, and civil rights) and argues that his results are consistent with the theory. See also Kessler et al. (1996) and Waldfogel's (1998) survey of the literature.

Although most of the literature on the selection of cases for trial uses the divergent expectations frameworks, some papers have investigated the relationship between trial rates and plaintiff win rates using asymmetric information. Froeb (1993) considers plea bargaining involving a privately informed defendant where guilty defendants are more likely to accept pleas. In his theoretical framework, factors that would increase the number of cases tried (such as a fall in prosecution costs) should increase the average guilt of defendants who go to trial. This pattern is documented in Froeb's sample of criminal cases.

3.9.3. Decoupling

The preceding discussion has assumed that any judgment against a defendant is automatically awarded to the plaintiff. Tying the defendant's liability to the plaintiff's award

[119] Many authors have documented systematic deviations from this number. See the references in Waldfogel (1998).

is also not necessarily optimal in theory. Polinsky and Che (1991) present a simple framework where a defendant chooses his level of precautions to reduce the probability of an accident, and an injured plaintiff would bring suit only if his or her expected award, a, exceeded the cost of litigating suit, c_p (which is distributed according to density $f(c_p)$ in the plaintiff population). Suppose the defendant's liability is l. There are two potential sources of inefficiency in this framework: (1) inefficient precautions by the defendant and (2) the wasted resources associated with litigation. The first-best outcome is not achievable when the defendant's liability is tied to the plaintiff's award. The award/liability level that achieves zero process costs, $a = l = 0$, provides the defendant with no incentives for care. (Conversely, the level that provides incentives for care leads to a deadweight loss at trial.) The optimal decoupled scheme makes the plaintiff's award, a, very small so that only a handful of cases are brought, and, at the same time, it makes the defendant's liability l very large so that his expected future liability equals the social harm that his actions cause. Decoupling creates a strong incentive for settlement because it creates a wedge between the most that the defendant is willing to pay and the least that the plaintiff is willing to accept: $[a - c_p, l + c_d]$.

Decoupling liability also appears in practice. Several states, including Iowa and Indiana, have adopted split-award statutes where the government keeps a percentage of punitive awards.[120] Kahan and Tuckman (1995) observe that these statutes can lead to an uneven playing field since the defendant's stakes are so much larger than the plaintiff's. Consequently, the defendant's incentives for care may be compromised by the statute because the plaintiff's incentives to invest in the case following an accident are reduced. See also Choi and Sanchirico (2004). Incentives are further compromised because the negotiated settlement will reflect, to a greater or lesser extent, the plaintiff's award at the bottom of the bargaining zone. See Daughety and Reinganum (2003).

3.9.4. Patent litigation

Suppose that a patentee and an imitator are negotiating prior to trial. If the case goes before a judge, there is a chance that the patent will be invalidated and the imitator will subsequently compete on equal footing with the innovator. The private settlement of these cases can be a legal way for competitors (or potential competitors) to collude and preserve market power at the expense of consumers.

Collusion through settlement is perhaps most obvious when the settlement between an owner of a patent and a current (or potential) imitator involves a lump-sum fee along with an agreement not to compete in the future. The patentee and the imitator can divide the monopoly profits between them through the lump-sum payment according to their relative bargaining strengths. Collusion can be achieved in more subtle ways, including

[120] In Iowa, for example, the plaintiff keeps only 25% of the punitive damage award. This example and others are discussed in Daughety and Reinganum (2003). See Landeo, Nikitin, and Babcock (2007) for a recent experimental analysis.

joint ventures (where one supplies the other) and a variety of royalty arrangements. It is not surprising that the Department of Justice and the Federal Trade Commission have watched these settlements with interest and have investigated many of them. See Shapiro (2003) for a good discussion of these mechanisms, the relevant cases, and proposed criteria for judicial approval of patent settlements.[121]

Competitors may also be able to use the threat of patent litigation to soften competition in a patent race. Marshall, Meurer and Richard (1994) describe a scenario where the loser of a patent race retains the option of bringing an action against the winner of the race.[122] Since the loser may succeed in this action, the possibility of patent litigation serves to reduce the differential value between winning the race and losing the race. The competitors subsequently spend less on research than otherwise, something that is typically in their joint interest. The overall impact of this collusive strategy on society more broadly (which includes consumer welfare) could be either positive or negative. See Reinganum (1989) for an excellent survey of the patent race literature.

Choi (1998) explores strategic decision making and externalities in repeated patent litigation. When a patentee is deciding whether or not to pursue an imitator, he must think carefully about the impact that a legal decision will have on future entrants. If the patent is deemed invalid and subsequently revoked, then the floodgates will open and industry profits will be dissipated. (On the other hand, if the patent is deemed valid, then the patentee will enjoy greater protection from the court's decision.) The patentee's decision to litigate the case depends upon the nature of these externalities.

Choi shows that when the patent is very likely to be upheld, then the patentee has an incentive to litigate. Less obviously, when the patent is likely to be overturned, then the patentee still has an incentive to litigate. The upside, the off-chance that the current imitator will be eliminated from the market, is stronger than the downside, the increase in entry once the patent is invalidated. The key to this last result is that entry will occur with or without a formal invalidation when the patent is very weak to begin with. When the probability of patent invalidation is in an intermediate range, however, the patentee chooses to tolerate early infringement and imitation.

Lanjouw and Schankerman (2001) document interesting correlations between litigation decisions and the characteristics of the patents that are involved. Many of the empirical findings are consistent with the predictions of the theory. First, they show that a patent is more likely to be litigated if it serves as the "base of a cumulative chain"—in other words, there are more rents to captured from future innovators. Second, they

[121] Meurer (1989) argues that patent cases may lead to more rather than less litigation when the settlements are closely regulated. To take an extreme case, when a regulator can force the patentee and the imitator to compete duopoly style following a settlement, it would be in their joint interest to go to trial in the hopes of establishing patent validity and the chance of monopoly rents.

[122] The patent example is informally described in the conclusion of their paper. Their formal model is one of procurement, where the loser of the auction can take action to reverse the outcome. The possibility of future legal action softens competition between the bidders during the auction. Intuitively, the differential value between winning and losing is smaller than before, since the loser may succeed in winning upon protest or appeal.

find support for the idea that firms establish reputations for protecting their intellectual property through litigation.

3.9.5. Insurance contracts

It is common for liability insurance contracts to: (1) place a dollar limit on coverage and (2) impose on the insurer a duty to defend the case, including the authority over settlement and the obligation to bear the legal costs of defense. We will see that these contracts create a conflict of interest between the insurer and the holder of the policy in the event of an accident, that these contracts potentially impose negative externalities on the victims of accidents, and that there is scope for legal intervention in these contracts.

Suppose that a victim—the plaintiff—has been harmed by an insured defendant. The defendant has an insurance policy, provided by a competitive insurance market, with a coverage limit of L. The plaintiff expects a random return at trial, x, from distribution $f(x)$ with expectation $E(x)$. If the case goes to trial and $x > L$ then the insurance company pays L and the defendant bears the residual, $x - L$. Suppose further that the insurer and the plaintiff would each bear litigation costs c_p and c_d if the case went to trial. (This is consistent with the insurance company's so-called "duty to defend.") If the case went to trial, the plaintiff would receive $E(x) - c_p$ in expectation—the expected damages at trial minus her litigation costs. When taken jointly, the insurance company and the defendant expect to pay $E(x) + c_d$. If they were jointly negotiating with the plaintiff, the case would surely settle for some amount in this range.

Conflicts of interest between the defendant and the insurer arise when the insurance company holds the authority to settle (Meurer, 1992; Sykes, 1994). If the case goes to trial, the insurer's expected losses are:

$$\int_0^L xf(x)dx + \int_L^\infty Lf(x)dx + c_d < E(x) + c_d.$$

The coverage limit makes the insurer tough in negotiations, lowering the most that he is willing to pay in settlement. This is because the defendant bears the downside associated with adverse judgments above the policy limit: $\int_L^\infty (x - L)f(x)dx$. Consequently, the insurance company may choose to litigate the case even though it is in the joint interest of the insurance company and the defendant to settle the case, instead.

Why would the defendant and the insurer ever write a contract that creates these *ex post* conflicts? Meurer (1992) argues that the insurance contract has strategic value for the defendant: it is a strategic commitment to be "tough" in settlement negotiations. By reducing the most that the insurer is willing to pay in settlement, the contract serves to extract value from the plaintiff by lowering the plaintiff's settlement demand. To put it somewhat differently, even though the conflict of interest may end up harming the defendant *ex post*, it may be in his advantage to sign such a contract *ex ante*. Indeed, a risk-neutral defendant with no liquidity constraints has an incentive to buy an insurance contract with these features. The value is coming not from risk aversion but from the strategic effects in settlement negotiations.

These insurance contracts are problematic from a social welfare perspective when neither the defendant nor the insurance company can foresee the distribution $f(x)$ at the time of contracting. Settlement negotiations can break down on the equilibrium path in this instance, leading to the *dead-weight loss* of trials $(c_p + c_d)$. In theory, this is similar to Aghion and Bolton's (1987) analysis of liquidated damages clauses. There, supply contracts with penalty clauses for breach have private strategic value because they encourage entrants to price more aggressively but create a social loss when they prevent entry altogether. Here, the insurance contract induces the plaintiff to accept a lower settlement amount but may lead to a breakdown in negotiations altogether.[123]

The *ex post* conflicts of interest between defendants and their liability insurers are mitigated in practice by the duty of liability insurers to act in good faith. A famous case, *Crisci v. Security Insurance Co.*, laid out the obligations of insurers in litigating and settling cases.[124] In particular, the court held that "the insurer must give the interests of the insured at least as much consideration as it gives its own interests."[125] Formally, this may be interpreted as forcing the insurer to equally weigh the defendant's expected payments at trial, $\int_L^\infty (x - L) f(x) dx$. This would be in the interest of public policy, as well. Crisci removes the insurer's incentive to play tough in negotiations and achieve a settlement in the range $[E(x) - c_p, E(x) + c_d]$.

3.9.6. *Financial distress*

Spier (2002) considers the role of settlement externalities in negotiations between one defendant and two plaintiffs when the defendant's wealth is insufficient to cover the total level of damages should both plaintiffs win at trial. Negotiations can break down when the probabilities that the plaintiffs will win at trial are sufficiently positively correlated. This may be seen in a simple example. Suppose that two plaintiffs, each of whom has suffered damages of \$80, are suing the same defendant. The defendant's wealth is \$80, sufficient to cover exactly one claim. Suppose that each plaintiff will prevail with probability $1/2$ and that the claims are perfectly correlated. If both plaintiffs go to trial, they each will receive \$20 in expectation (they win with probability $1/2$ and then split the defendant's wealth between them). But the plaintiffs will refuse to settle for \$20 each: if one plaintiff accepts the \$20, the other would rather go to trial with the expectation of winning the remaining \$60 with probability $1/2$! When the plaintiffs are acting individually, the defendant would have to offer a hefty settlement premium to get both to accept.

Suppose instead that the plaintiffs' cases are uncorrelated. The expected return for each plaintiff is \$30 if both cases go to trial. Perhaps surprisingly, the defendant may

[123] Note that the social cost of trials may be avoidable here if the defendant can engage in negotiations as well and contribute to the settlement. In practice, transactions costs and information asymmetries may interfere with renegotiation along these lines. See Sykes (1994).

[124] 66 Cal. 2d 425, 58 Cal Rptr. 13, 426 P.2d 173 (1967).

[125] 66 Cal. 2d at 429.

succeed in coercing the plaintiff to settle for less than this amount. If one plaintiff settles out of court for $30, then $50 remains in the defendant's coffers. Since the non-settling plaintiff wins at trial half the time, his expected return is only $25. In general, when the degree of correlation between the cases is low, then the defendant's insolvency creates a "rush to collect" by the plaintiffs and may allow the defendant to successfully chisel his offers.

Spier and Sykes (1998) explore how the capital structure of a defendant can affect negotiations with a single plaintiff when a civil judgment may bankrupt the firm. Not surprisingly, the outcome of settlement negotiations depends critically on a firm's debt level and on the priority afforded the debtholders in bankruptcy. Most interestingly, they show that even junior debt can have strategic value by making the shareholders into tougher negotiators. The intuition rests on the fact that while any settlement that is paid will come out of the shareholders' pockets, a large award at trial will (eventually) come out of the debtholders' pockets. Although the presence of junior debt does not directly dilute the value of the tort claim, it does so indirectly through its impact on the bargaining range.

3.9.7. Strict liability and negligence rules

There is a large law and economics literature that compares strict liability to negligence.[126] Under strict liability, a defendant is held responsible for a plaintiff's damages regardless of his level of care—the defendant is forced to compensate the plaintiff even if he took appropriate precautions to avoid the accident. Under the negligence rule, the defendant is held liable only if he failed to take due care. At first glance, strict liability appears to have higher transactions costs because every single accident involving the defendant creates a meritorious legal claim. Under the negligence rule, on the other hand, a meritorious claim arises only if the defendant was negligent. There are countervailing factors, however. First, a given case brought under a negligence rule may be more expensive to try: the plaintiff must establish the defendant's negligence in addition to causation. Second, cases brought under a negligence rule may be less likely to settle since there is more scope for disagreement because the plaintiff and defendant may have different perceptions or information concerning the defendant's level of care. Again, this would lead to higher transactions costs than otherwise. Finally, the negligence rule can lead the defendant to engage in higher levels of the risky economic activity, creating more accidents and potentially more lawsuits. When these additional factors are taken into account, the negligence rule may actually be more expensive than strict liability.

[126] See Shavell (2004) for a survey of this literature. In a frictionless world, both rules can lead to socially optimal levels of care although they differ in the effects on the level of economic activity.

3.9.8. Scheduled damages

Another strand of the literature focuses on the accuracy of legal rules. In a model without settlement, Kaplow and Shavell (1994, 1996) argued that legal rules that fine-tune liability to the plaintiff's actual level of harm lead to social waste. It is expensive for the litigants (and for the court) to verify the precise level of damages at trial. The "scheduling" of damages, or standardizing awards for injuries that fall into particular categories (as in workers' compensation), can reduce the transactions costs of litigation. Scheduling can be valuable when settlement is possible as well (Spier, 1994c). If the plaintiff has private information about his damages, for example, then negotiations will sometimes break down and litigation costs will be borne. Scheduling makes the future outcome of the case more transparent—there is less to argue about—and can help to promote settlement and save on litigation costs. There may be a downside of damage schedules, however: schedules may be less effective in encouraging potential injurers to take precautions to reduce the magnitude of harm. Spier (1992b) makes a related point in the context of contract design. In a risk-sharing model, it is shown that the benefits associated with optimal risk sharing may be outweighed by the transactions costs of enforcing these contracts (e.g., costly state verification and equilibrium litigation activity).

4. Conclusion

The purpose of this chapter has been to survey the vast literature on litigation and settlement and to synthesize its main themes. Although the set of issues raised in this chapter are very broad and far reaching, the approach taken in this chapter was quite focused. The first underlying premise was that those engaged in litigation—the litigants, their attorneys, the judges, and other interested parties—are rational, self-interested, and strategic. A second premise was that the primary purpose of the court system should be to facilitate value-creating economic activity and deter value-destroying economic activity.

Section 2 of this chapter laid out the basic economic framework of a potential lawsuit involving a single plaintiff and a single defendant. It carefully examined the plaintiff's decision to pursue litigation and the decision of the plaintiff and defendant to settle their case out of court. The positive and normative implications that arise in this setting are surprisingly subtle. We saw that the plaintiff and defendant's litigation strategies depend critically upon the timing of moves and on the information available to them. We also saw that the private interests of the plaintiff and the defendant in litigation are not necessarily aligned with the interests of society, more broadly. The plaintiff may bring "too many" lawsuits, choosing to hire an expensive attorney to battle a defendant in court even in situations when the transfer of funds has little corresponding social benefit. The plaintiff may also pursue "too few" lawsuits, particularly if the anticipation of legal action is a public good, giving potential defendants incentives to take precautions.

Since an individual plaintiff does not personally capture the social value of deterrence in this instance, she may not sue often enough, from a societal perspective. Similarly, the level of expenditures by the plaintiff and the defendant and their private incentives to settle their dispute are not necessarily aligned with the interests of society. Section 3 showed how this framework has been productively applied in the law and economics literature.

Economic approaches to studying litigation will likely continue to be popular in the future. It is clear that a deeper understanding of the economic issues is necessary and critical for those who are creating public policy. The subtleties—and excitement—of these issues will also continue to capture the imagination of talented young economists and legal scholars for years to come. One way that the frontiers are being pushed forward is through the introduction of new methods and approaches. The law and economics field more broadly has experienced a growing influence of behavioral assumptions and experimental techniques in addition to the ongoing growth of empirical work. These approaches are very valuable, allowing us to resolve theoretical ambiguity and pinpoint the magnitude of economic effects. On the theoretical side, there is a definite need to explore more fully the normative implications of settlement and litigation. There has been relatively little work, for example, on how the transactions costs of litigation affect the design of contracts, the organization of economic activity, and the boundaries of the firm.

References

Aghion, P., Bolton, P. (1987). “Contracts as a barrier to entry”. American Economic Review 77, 388–401.

Andreoni, J. (1991). “Reasonable doubt and the optimal magnitude of fines: should the penalty fit the crime?” RAND Journal of Economics 22, 385–395.

Ashenfelter, O., Bloom, L. (1984). “Models of arbitrator behavior: Theory and evidence”. American Economic Review 74, 111–124.

Ashenfelter, O., Currie, J., Farber, H.S., Spiegel, M. (1992). “An experimental comparison of dispute rates in alternative arbitration systems”. Econometrica 6, 1407–1433.

Aumann, R. (1976). “Agreeing to disagree”. Annals of Statistics 4, 1236–1239.

Austen-Smith, D., Banks, J. (1996). “Information aggregation, rationality, and the Condorcet jury theorem”. American Political Science Review 90, 34–45.

Ayres, I., Nalebuff, B. (1997). “Common knowledge as a barrier to negotiation”. UCLA Law Review 80, 323–401.

Baker, S., Mezzetti, C. (2001). “Prosecutorial resources, plea bargaining and the decision to go to trial”. Journal of Law, Economics & Organization 17, 149–167.

Bar-Gill, O. (2006). “Evolution and persistence of optimism in litigation”. Journal of Law, Economics & Organization 22, 490–507.

Bebchuk, L.A. (1984). “Litigation and settlement under imperfect information”. RAND Journal of Economics 15, 404–415.

Bebchuk, L.A. (1988). “Suing solely to extract a settlement offer”. Journal of Legal Studies 17, 437–450.

Bebchuk, L.A. (1996). “A new theory concerning the credibility and success of threats to sue”. Journal of Legal Studies 25, 1–25.

Bebchuk, L.A., Chang, H. (1996). “An analysis of fee shifting based on the margin of victory: On frivolous suits, meritorious suits, and the role of Rule 11”. Journal of Legal Studies 25, 371–403.

Bebchuk, L.A., Chang, H. (1999). “The effect of offer-of-settlement rules on the terms of settlement”. Journal of Legal Studies 28, 489–513.

Bebchuk, L.A., Guzman, A. (1996). “How would you like to pay for that? The strategic effects of fee arrangements on settlement terms”. Harvard Negotiation Law Review 1, 53–63.

Bebchuk, L.A., Kaplow, L. (1992). “Optimal sanctions when individuals are imperfectly informed about the probability of apprehension”. Journal of Legal Studies 21, 365–370.

Becker, G. (1968). “Crime and punishment: And economic approach”. Journal of Political Economy 76, 169–217.

Bernardo, A.E., Talley, E., Welch, I. (2000). “A theory of legal presumptions”. Journal of Law, Economics, & Organization 16, 1–49.

Bernstein, L. (1992). “Opting out of the legal system: Extralegal contractual relations in the diamond industry”. Journal of Legal Studies 21, 115–157.

Bernstein, L. (1993). “Understanding the limits of court-connected ADR: A critique of federal court-annexed arbitration programs”. University of Pennsylvania Law Review 141, 2169–2233.

Bikchandani, S., Hirschleifer, D., Welch, I. (1998). “Learning from the behavior of others: Conformity, fads, and informational cascades”. Journal of Economic Perspectives 12, 151–170.

Borenstein, S. (1996). “Settling for coupons: Discount contracts as compensation and punishment in antitrust lawsuits”. Journal of Law and Economics 39, 379–404.

Braeutigam, R., Owen, B., Panzar, J. (1984). “An economic analysis of alternative fee shifting systems”. Law and Contemporary Problems 47, 173–204.

Bulow, J.I., Klemperer, P.D. (1998). “The tobacco deal”. Brookings Papers on Economic Activity: Microeconomics, 323–394.

Bulow, J.I., Geanakoplos, J.D., Klemperer, P.D. (1985). “Multimarket oligopoly: Strategic substitutes and complements”. Journal of Political Economy 93, 488–511.

Bustos, A.E. (2006). A Dynamic Theory of Corporate Common Law, Princeton University Doctoral Dissertation.

Butz, D.A. (1990). “Durable goods monopoly and best-price provisions”. American Economic Review 80, 1062–1076.

Chang, H.F., Sigman, H. (2000). “Incentives to settle under joint and several liability: An empirical analysis of superfund litigation”. Journal of Legal Studies 24, 205–236.

Che, Y.-K. (1996). “Equilibrium formation of class action suits”. Journal of Public Economics 62, 339–361.

Che, Y.-K. (2002). “The economics of collective negotiations in pretrial bargaining”. The International Economic Review 43, 549–576.

Che, Y.-K., Yi, J.G. (1993). “The role of precedents in repeated litigation”. Journal of Law, Economics, and Organization 9, 399–424.

Chen, K.-P., Chien, H.-K., Chu, C.Y.C. (1997). “Sequential versus unitary trials with asymmetric information”. The Journal of Legal Studies 26, 239–258.

Cho, I.-K., Kreps, D. (1987). “Signalling games and stable equilibria”. Quarterly Journal of Economics 102, 179–221.

Choi, J.P. (1998). “Patent litigation as an information-transmission mechanism”. American Economic Review 88, 1249–1263.

Choi, A. (2003). “Allocating settlement authority under a contingent-fee arrangement”. The Journal of Legal Studies 32, 585–610.

Choi, A., Sanchirico, C.W. (2004). “Should plaintiffs win what defendants lose? Litigation stakes, litigation effort, and the benefits of decoupling”. Journal of Legal Studies 22, 323–354.

Clermont, K., Eisenberg, T. (1992). “Trial by jury or judge: Transcending empiricism”. Cornell University Law Review 77, 1124–1177.

Condorcet, M. de (1785). Essais Sur L’appication de L’analyse a la Probabilite des Decisions Rendues a la Puralite des Voix, Paris (Trans. 1994 by Iain McLean and Fiona Hewitt).

Cooter, R., Kornhauser, L., Lane, D. (1979). “Liability rules, limited information, and the role of precedent”. The Bell Journal of Economics 10, 366–373.

Cooter, R.D., Rubinfeld, D.L. (1989). “Economic analysis of legal disputes and their resolution”. Journal of Economic Literature 27, 1067–1097.

Cooter, R.D., Rubinfeld, D.L. (1994). “An economic model of legal discovery”. Journal of Legal Studies 23, 435–464.

Dana, J.D., Spier, K.E. (1993). “Expertise and contingent fees: The role of asymmetric information in attorney compensation”. The Journal of Law, Economics, & Organization 9, 349–367.

Danzon, P.M. (1983). “Contingent fees for personal injury litigation”. Bell Journal of Economics 14, 213–223.

Danzon, P., Lillard, L. (1983). “Settlement out of court: The disposition of medical Malpractice Claims”. Journal of Legal Studies 12, 345–378.

Daughety, A.F. (2000). “Settlement”. In: Bouckaert, B., De Geest, G. (Eds.), Encyclopedia of Law and Economics, vol. 5. Edward Elgar Publishing Co, Cheltenham, pp. 95–158.

Daughety, A.F., Reinganum, J.F. (1993). “Endogenous sequencing in models of settlement and litigation”. Journal of Law, Economics and Organization 9, 314–349.

Daughety, A.F., Reinganum, J.F. (1994). “Settlement negotiations with two-sided asymmetric information: Model duality, information distribution and efficiency”. International Review of Law and Economics 14, 283–298.

Daughety, A.F., Reinganum, J.F. (1995). “Keeping society in the dark: On the admissibility of pretrial negotiations as evidence in court”. RAND Journal of Economics 26, 203–221.

Daughety, A.F., Reinganum, J.F. (1999a). “Hush money”. RAND Journal of Economics 30, 661–678.

Daughety, A.F., Reinganum, J.F. (1999b). “Stampede to judgment: Persuasive influence and herding behavior by courts”. American Law and Economics Review 1, 158–189.

Daughety, A.F., Reinganum, J.F. (2000a). “Appealing judgments”. RAND Journal of Economics 31, 502–525.

Daughety, A.F., Reinganum, J.F. (2000b). “On the economics of trials: Adversarial process, evidence, and equilibrium bias”. Journal of Law, Economics, & Organization 16, 365–394.

Daughety, A.F., Reinganum, J.F. (2002). “Information externalities in settlement bargaining: Confidentiality and correlated culpability”. RAND Journal of Economics 33 (4), 587–604.

Daughety, A.F., Reinganum, J.F. (2003). “Found money? Split-award statutes and settlement of punitive damages cases”. American Law and Economics Review 5, 134–164.

Daughety, A.F., Reinganum, J.F. (2004). “Exploiting future settlements: A signaling model of most-favored-nation clauses in settlement bargaining”. RAND Journal of Economics 35 (3), 467–485.

Daughety, A.F., Reinganum, J.F. (2005). “Secrecy and safety”. American Economic Review 95, 1074–1091.

Demougin, D., Fluet, C. (2006). “Preponderance of evidence”. European Economic Review 50, 963–976.

Dewatripont, M., Tirole, J. (1999). “Advocates”. Journal of Political Economy 107, 1–39.

Dixit, A. (1987). “Strategic behavior in contests”. American Economic Review 77, 891–898.

Dixit, A. (2003). “Arbitration and information”, Princeton University Working Paper.

Emons, W. (2000). “Expertise, contingent fees, and insufficient attorney effort”. International Review of Law and Economics 20, 21–33.

Farber, H.S. (1980). “An analysis of final offer arbitration”. Journal of Conflict Resolution 35, 683–705.

Farber, H.S., White, M.J. (1991). “Medical malpractice: An empirical examination of the litigation process”. Rand Journal of Economics 22, 199–217.

Farmer, A., Pecorino, P. (1994). “Pretrial negotiations with asymmetric information on risk preferences”. International Review of Law and Economics 14, 273–281.

Farmer, A., Pecorino, P. (2000). “Conditional cost shifting and the incidence of trial: Pretrial bargaining in the face of a Rule 68 offer”. American Law and Economics Review 2, 318–340.

Farmer, A., Pecorino, P. (2003). “Bargaining with voluntary transmission of private information: Does the use of final offer arbitration impede settlement?” Journal of Law, Economics & Organization 19, 64–82.

Feddersen, T., Pesendorfer, W. (1998). “Convicting the innocent: The inferiority of unanimous jury verdicts”. American Political Science Review 92, 23–35.

Fon, V., Parisi, F. (2003). “Litigation and the evolution of legal remedies: A dynamic analysis”. Public Choice 116, 419–433.

Fon, V., Parisi, F. (2004). “Judicial precedents in civil law systems: A dynamic analysis”, George Mason Law & Economics Research Paper No. 04-15.

Fournier, G.M., Zuehlke, T.W. (1996). "The timing of out-of-court settlements". RAND Journal of Economics 27, 310–321.

Froeb, L.M. (1993). "Adverse selection of cases for trial". International Review of Law and Economics 13, 317–324.

Froeb, L.M., Kobayashi, B.H. (1996). "Naive, biased, yet bayesian: Can juries interpret selectively produced evidence?". Journal of Law, Economics, & Organization 12, 257–276.

Gennaioli, N., Shleifer, A. (2005). "The evolution of precedent", NBER Working Paper No. W11265.

Gibbons, R. (1988). "Learning in equilibrium models of arbitration". American Economic Review 78, 896–912.

Gould, J. (1973). "The economics of legal conflicts". Journal of Legal Studies 2, 279–300.

Grossman, S.J. (1981). "The informational role of warranties and private disclosure about product quality". Journal of Law and Economics 24, 461–483.

Grossman, G.M., Katz, M.L. (1983). "Plea bargaining and social welfare". American Economic Review 73, 749–757.

Hadfield, G.K. (2006). "The quality of law in civil code and common law regimes: Judicial incentives, legal human capital, and the evolution of law", working paper. (Available at http://law.bepress.com/alea/16th/art40.)

Hause, J. (1989). "Indemnity, settlement, and litigation, or I'll be suing you". Journal of Legal Studies 18, 157–180.

Hay, B.L. (1993). "Some settlement effects of preclusion". Illinois Law Review 1993, 21–57.

Hay, B.L. (1994). "Civil discovery: Its effects and optimal scope". Journal of Legal Studies 23, 481–517.

Hay, B.L. (1995). "Effort, information, settlement, trial". Journal of Legal Studies 24, 29–62.

Hay, B.L. (1997). "Optimal contingent fees in a world of settlement". Journal of Legal Studies 26, 259–278.

Hay, B.L., Spier, K.E. (1997). "Burdens of proof in civil litigation: An economic perspective". Journal of Legal Studies 26, 413–433.

Hay, B., Spier, K.E. (1998). "Settlement of litigation". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 3. Macmillan Reference Limited, London, pp. 442–451.

Helland, E., Tabarrok, A. (2000). "Runaway judges? Selection effects and the jury". Journal of Law, Economics & Organization 16, 306–333.

Helland, E., Tabarrok, A. (2003). "Contingency fees, settlement delay, and low-quality litigation: Empirical evidence from two datasets". Journal of Law, Economics, and Organization 19, 517–542.

Hua, X., Spier, K.E. (2005). "Information and externalities in sequential litigation". Journal of Institutional and Theoretical Economics 161, 215–232.

Hughes, J.W., Snyder, E.A. (1998). "Allocation of legal costs: American and English rules". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 1. Macmillan Reference Limited, London, pp. 51–56.

Hylton, K.N. (2002). "Welfare implication of costly litigation under strict liability". American Law and Economics Review 4, 18–42.

Hylton, K.N. (2006). "Information, litigation, and common law evolution". American Law and Economics Review 8, 33–61.

Jost, P.J. (1995). "Disclosure of information and incentives for care". International Review of Law and Economics 15, 65–85.

Judicial Business of the United States Courts (2001). Annual Report 2001. Available online at: http://www.uscourts.gov/judbus2001/contents.html.

Kalven, H. Jr., Zeisel, H. (1966). The American Jury. Little Brown, Boston.

Kahan, M., Tuckman, B. (1995). "Special levies for punitive damages". International Review of Law and Economics 15, 175–185.

Kaplow, L. (1993). "Shifting plaintiffs' fees versus increasing damage awards". RAND Journal of Economics 24, 625–630.

Kaplow, L. (1998). "Accuracy in adjudication". In: The New Palgrave Dictionary of Economics and the Law, vol. 1. Macmillan Reference Limited, London, pp. 1–7.

Kaplow, L., Shavell, S. (1992). "Private versus socially optimal provision of ex-ante legal advice". Journal of Law, Economics, and Organization 8, 306–320.

Kaplow, L., Shavell, S. (1994). "Accuracy in the assessment of liability". Journal of Law and Economics 37, 1–16.

Kaplow, L., Shavell, S. (1996). "Accuracy in the assessment of damages". Journal of Law and Economics 39, 191–209.

Katz, A.W. (1987). "Measuring the demand for litigation: Is the English Rule really cheaper?". Journal of Law, Economics, and Organization 3, 143–176.

Katz, A.W. (1988). "Judicial decisionmaking and litigation expenditure". International Review of Law and Economics 8, 127–143.

Katz, A.W. (1990). "The effect of frivolous lawsuits on the settlement of litigation". International Review of Law and Economics 10, 3–27.

Kim, J.-Y., Ryu, K. (2000). "Pretrial negotiation behind closed doors versus open doors: Economic analysis of Rule 408". International Review of Law and Economics 20, 285–294.

Kessler, D., Meites, T., Miller, G.P. (1996). "Explaining deviations from the 50% rule: A multimodal approach to the selection of cases for litigation". Journal of Legal Studies 25, 233–259.

Klerman, D. (1996). "Settling multidefendant lawsuits: The advantage of conditional setoff rules". Journal of Legal Studies 25, 445–461.

Kornhauser, L.A. (1992). "Modeling collegial courts, 2: Legal doctrine". Journal of Law, Economics, and Organization 8, 441–470.

Kornhauser, L.A., Revesz, R.L. (1994a). "Multidefendant settlements: The impact of joint and several liability". Journal of Legal Studies 23, 41–76.

Kornhauser, L.A., Revesz, R.L. (1994b). "Multidefendant settlements under joint and several liability: The problem of insolvency". Journal of Legal Studies 23, 517–542.

Kornhauser, L.A., Revesz, R.L. (1998). "Joint and several liability". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 2. Macmillan Reference Limited, London, pp. 371–376.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R. (1997). "Legal determinants of external finance". Journal of Finance 52, 1131–1150.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R. (1998). "Law and finance". Journal of Political Economy 106, 1113–1155.

Landeo, C.M., Nikitin, M., Babcock, L. (2007). "Split-awards and disputes: An experimental study of a strategic model of litigation". Journal of Economic Behavior and Organization 63, 553–572.

Landes, W.M. (1971). "An economic analysis of the courts". Journal of Law and Economics 14, 61–107.

Landes, W.M. (1993). "Sequential versus unitary trials: An economic analysis". Journal of Legal Studies 22, 99–134.

Landes, W.M., Posner, R.A. (1976). "Legal precedents: A theoretical and empirical analysis". Journal of Law and Economics 19, 249–307.

Lanjouw, J.O., Schankerman, M. (2001). "Characteristics of patent litigation: A window on competition". RAND Journal of Economics 32, 129–151.

Levy, G. (2005). "Careerist judges". RAND Journal of Economics 36, 275–297.

Lewis, T., Poitevin, M. (1997). "Disclosure of information in regulatory proceedings". Journal of Law, Economics & Organization 13, 50–73.

Loewenstein, G., et al. (1993). "Self-serving assessments of fairness and pretrial bargaining". Journal of Legal Studies 22, 135–158.

Marshall, R.C., Meurer, M.J., Richard, J.-F. (1994). "Litigation settlement and collusion". Quarterly Journal of Economics 109, 211–239.

Meurer, M. (1989). "The settlement of patent litigation". RAND Journal of Economics 20, 77–91.

Meurer, M. (1992). "The gains from faith in an unfaithful agent: Settlement conflict between defendants and liability insurer". Journal of Law, Economics and Organization 8, 502–522.

Miceli, T. (1994). "Do contingency fees promote excessive litigation?". Journal of Legal Studies 23, 211–224.

Miceli, T., Cosgel, M.M. (1994). "Reputation and judicial decision making". Journal of Economic Behavior and Organization 23, 31–51.

Milgrom, P., Roberts, J. (1986). “Relying on the information of interested parties”. RAND Journal of Economics 17, 18–32.

Milgrom, P., Weber, R. (1982). “A theory of auctions and competitive bidding”. Econometrica 50, 1089–1122.

Miller, G.P. (1986). “An economic analysis of Rule 68”. Journal of Legal Studies 15, 93–125.

Miller, G.P. (1987). “Some agency problems in settlement”. Journal of Legal Studies 16 (1), 189–215.

Miller, G.P. (1998). “Class Actions”. In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 1. Macmillan Reference Limited, London, pp. 257–262.

Mnookin, R.H. (1998). “Alternative dispute resolution”. In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 1. Macmillan Reference Limited, London, pp. 56–60.

Mnookin, R.H., Wilson, R. (1998). “A model of efficient discovery”. Games and Economic Behavior 25, 219–250.

Mnookin, R.H., Kornhauser, L. (1979). “Bargaining in the shadow of the law: The case of divorce”. Yale Law Journal 88, 950–997.

Mukhopadhaya, K. (2003). “Jury size and the free rider problem”. Journal of Law, Economics & Organization 19, 24–44.

Myerson, R. (1979). “Incentive compatibility and the bargaining problem”. Econometrica 47, 61–73.

Myerson, R., Satterthwaite, M. (1983). “Efficient mechanisms for bilateral trading”. Journal of Economic Theory 29, 265–281.

Nalebuff, B. (1987). “Credible pretrial negotiation”. RAND Journal of Economics 18, 198–210.

Neeman, Z., Klement, A. (2005). “Against compromise: A mechanism design approach”. Journal of Law, Economics, & Organization 21, 285–314.

Note (1992). “Exposing the extortion gap: An economic analysis of the rules of collateral estoppel”. Harvard Law Review 105, 1940–1960.

Ostrom, B.J., Kauder, N.B., LaFountain, R.C. (2001). Examining the Work of the State Courts, 1999–2000. National Center for State Courts, Williamsburg, VA.

Parisi, F. (2002). “Rent seeking through litigation: Adversarial and inquisitorial systems compared”. International Review of Law and Economics 22, 193–216.

P’ng, I.P.L. (1983). “Strategic behavior in suit, settlement, and trial”. RAND Journal of Economics 14, 539–550.

P’ng, I.P.L. (1986). “Optimal subsidies and damages in the presence of judicial error”. International Review of Law and Economics 6, 101–105.

Polinsky, A.M., Che, Y.-K. (1991). “Decoupling liability: Optimal incentives for care and litigation”. Rand Journal of Economics 22, 562–570.

Polinsky, A.M., Rubinfeld, D.L. (1988a). “The deterrent effects of settlements and trials”. International Review of Law and Economics 8, 109–116.

Polinsky, A.M., Rubinfeld, D.L. (1988b). “The welfare implications of costly litigation for the level of liability”. Journal of Legal Studies 17, 151–164.

Polinsky, A.M., Rubinfeld, D.L. (1996). “Optimal awards and penalties when the probability of prevailing varies among plaintiffs”. RAND Journal of Economics 27, 269–280.

Polinsky, A.M., Rubinfeld, D.L. (1998). “Does the English Rule discourage low-probability-of-prevailing plaintiffs?”. Journal of Legal Studies 27, 519–535.

Polinsky, A.M., Rubinfeld, D.L. (2003). “Aligning the interests of lawyers and clients”. American Law and Economics Review 5, 165–187.

Polinsky, A.M., Rubinfeld, D.L. (in preparation). “The deadweight loss of coupon remedies for price overcharges”. Journal of Industrial Economics.

Polinsky, A.M., Shavell, S. (1989). “Legal error, litigation, and the incentive to obey the law”. Journal of Law, Economics, and Organization 5, 99–108.

Posner, R.A. (1973). “An economic approach to legal procedure and judicial administration”. Journal of Legal Studies 2, 399–458.

Posner, R.A. (1992). Economic Analysis of Law, 4th edn. Little, Brown, Boston.

Posner, R.A. (2003). Law and Legal Theory in England and America (Clarendon Law Lecture). Oxford University Press, New York.

Priest, G. (1977). "Common law process and selection of efficient rules". Journal of Legal Studies 6, 65–82.

Priest, G., Klein, B. (1984). "The selection of disputes for litigation". Journal of Legal Studies 13, 1–55.

Rasmusen, E. (1994). "Judicial legitimacy as a repeated game". Journal of Law, Economics, and Organization 10, 63–83.

Reinganum, J.F. (1988). "Plea bargaining and prosecutorial discretion". American Economic Review 78, 713–728.

Reinganum, J. (1989). "The timing of innovation: Research, development, and diffusion". The Handbook of Industrial Organization 1, 849–908.

Reinganum, J.F. (2000). "Sentencing guidelines, judicial discretion, and plea bargaining". RAND Journal of Economics 31, 62–81.

Reinganum, J., Wilde, L. (1986). "Settlement, litigation, and the allocation of litigation costs". RAND Journal of Economics 17, 557–568.

Rickman, N. (1999). "Contingent fees and litigation settlement". International Review of Law and Economics 19, 295–317.

Rosenberg, D., Shavell, S. (1985). "A model in which lawsuits are brought for their nuisance value". International Review of Law and Economics 5, 3–13.

Rubin, P. (1977). "Why is the common law efficient?" Journal of Legal Studies 6, 51–63.

Rubinfeld, D.L., Sappington (1987). "Efficient awards and standards of proof in judicial proceedings". RAND Journal of Economics 18, 308–315.

Rubinfeld, D.L., Scotchmer, S. (1993). "Contingent fees for attorneys: An economic analysis". RAND Journal of Economics 24, 343–356.

Rubinfeld, D.L., Scotchmer, S. (1998). "Contingent fees". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 1. Macmillan Reference Limited, London, pp. 415–425.

Sanchirico, C.W. (1997). "The burden of proof in civil litigation: A simple model of mechanism design". International Review of Law and Economics 17, 431–447.

Sanchirico, C.W. (2001). "Relying on the information of interested—and potentially dishonest—parties". American Law and Economics Review 3, 320–357.

Santore, R., Viard, A.D. (2001). "Legal fee restrictions, moral hazard, and attorney rights". Journal of Law & Economics 44, 549–572.

Schrag, J. (1999). "Managerial judges: An economic analysis of the judicial management of legal discovery". RAND Journal of Economics 30, 305–323.

Schwartz, E. (1992). "Policy, precedent, and power: A positive theory of supreme-court decision making". Journal of Law, Economics, and Organization 8, 219–252.

Schwartz, W. (1967). "Severance—A means of minimizing the role of burden and expense in determining the outcome of litigation". Vanderbilt Law Review 20, 1197–1214.

Schwartz, M.L., Mitchell, D.J.B. (1970). "An economic analysis of the contingent fee and personal injury litigation". Stanford Law Review 22, 1125–1162.

Schweizer, U. (1989). "Litigation and settlement under two sided incomplete information". Review of Economic Studies 56, 163–177.

Shapiro, C. (2003). "Antitrust limits to patent settlements". RAND Journal of Economics 34, 391–411.

Shavell, S. (1982a). "Suit, settlement and trial: A theoretical analysis under alternative methods for the allocation of legal costs". Journal of Legal Studies 11, 55–82.

Shavell, S. (1982b). "The social versus the private incentive to bring suit in a costly legal system". Journal of Legal Studies 11, 333–339.

Shavell, S. (1989a). "Optimal sanctions and the incentive to provide evidence to legal tribunals". International Review of Law and Economics 9, 3–11.

Shavell, S. (1989b). "The sharing of information prior to settlement or litigation". RAND Journal of Economics 20, 183–195.

Shavell, S. (1993). "Suit versus settlement when parties seek nonmonetary judgments". Journal of Legal Studies 22, 1–14.

Shavell, S. (1994). "Alternative dispute resolution: An economic analysis". Journal of Legal Studies 24, 1–28.

Shavell, S. (1995). “The appeals process as a means of error correction”. Journal of Legal Studies 24, 379–426.

Shavell, S. (1996). “Any probability of plaintiff victory at trial is possible”. Journal of Legal Studies 25, 493–501.

Shavell, S. (1997). “The fundamental divergence between the private and the social motive to use the legal system”. Journal of Legal Studies 26, 575–613.

Shavell, S. (2004). Foundations of Economic Analysis of Law. Belknap Press of Harvard University Press, Cambridge.

Shepherd, G. (1999). “An empirical study of the effects of pretrial discovery”. International Review of Law and Economics 19, 245–263.

Shin, H.S. (1994). “The burden of proof in a game of persuasion”. Journal of Economic Theory 64, 253–264.

Shin, H.S. (1998). “Adversarial and inquisitorial procedures in arbitration”. Rand Journal of Economics 29, 378–405.

Sobel, J. (1985). “Disclosure of evidence and resolution of disputes: Who should bear the burden of proof?” In: Roth, A. (Ed.), Game Theoretic Models of Bargaining. Cambridge University Press, New York.

Sobel, J. (1989). “An analysis of discovery rules”. Law and Contemporary Problems 52, 133–159.

Spier, K.E. (1992a). “The dynamics of pretrial negotiation”. Review of Economic Studies 59, 93–108.

Spier, K.E. (1992b). “Incomplete contracts and signaling”. RAND Journal of Economics 23, 432–443.

Spier, K.E. (1994a). “A note on joint and several liability: Insolvency, settlement, and incentives”. Journal of Legal Studies 23, 559–568.

Spier, K.E. (1994b). “Pretrial bargaining and the design of fee-shifting rules”. RAND Journal of Economics 25, 197–214.

Spier, K.E. (1994c). “Settlement bargaining and the design of damage awards”. Journal of Law, Economics, & Organization 10, 84–95.

Spier, K.E. (1997). “A note on the divergence between the private and social motive to settle under a negligence rule”. Journal of Legal Studies 26, 613–623.

Spier, K.E. (2002). “Settlement with multiple plaintiffs: The role of insolvency”. Journal of Law, Economics, & Organization 18, 295–323.

Spier, K.E. (2003a). “The use of most-favored-nation clauses in settlement of litigation”. The RAND Journal of Economics 34, 78–95.

Spier, K.E. (2003b). “Tied to the mast: Most-favored-nation clauses in settlement contracts”. Journal of Legal Studies 32, 91–120.

Spier, K.E., Sykes, A.O. (1998). “Capital structure, priority rules, and the settlement of civil claims”. The International Review of Law and Economics 18, 187–200.

Spitzer, M., Talley, E. (2000). “Judicial auditing”. Journal of Legal Studies 24, 649–683.

Spurr, S.J. (1991). “An economic analysis of collateral estoppel”. International Review of Law and Economics 11, 47–61.

Statistical Abstract of the United States (2001). U.S. Department of Commerce, Economics and Statistics Administration, US Census Bureau, Washington, D.C.

Sykes, A.O. (1994). “Bad faith’ refusal to settle by liability insurers: Some implications of the judgment-proof problem”. Journal of Legal Studies 23, 77–110.

The 2001 Survey of Law Firm Economics (2001). Altman Weil, Inc., Altman Weil Publications, Newton Sq., Pa.

Thomason, T. (1991). “Are attorneys paid what they are worth? Contingency fees and the settlement process”. Journal of Legal Studies 20, 187–233.

Tullock, G. (1980). Trials on Trial. Columbia University Press, New York.

Waldfogel, J. (1995). “The selection hypothesis and the relationship between trial and plaintiff victory”. Journal of Political Economy 103, 229–260.

Waldfogel, J. (1998). “Selection of cases for trial”. In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 3. Macmillan Reference Limited, London, pp. 419–424.

Watanabe, Y. (2005). Learning and Bargaining in Dispute Resolution: Theory and Evidence from Medical Malpractice Litigation. Northwestern University mimeo.

White, M. (2002). “Explaining the flood of asbestos litigation: Consolidation, bifurcation, and bouquet trials”, National Bureau of Economic Research Working Paper, W9362.

Wittman, D. (2003). “Lay juries, professional arbitrators, and the arbitrator selection hypothesis”. American Law and Economics Review 5, 61–93.

Yildiz, M. (2003). “Bargaining without a common prior—an immediate agreement theorem”. Econometrica 71, 793–811.

Yildiz, M. (2004). “Waiting to persuade”. Quarterly Journal of Economics 119, 223–248.

Yoon, A. (2004). “Mandatory arbitration and civil litigation: An empirical study of medical malpractice litigation in the west”. American Law and Economics Review 6, 95–134.

Chapter 5

EMPIRICAL STUDY OF THE CIVIL JUSTICE SYSTEM

DANIEL P. KESSLER

Graduate School of Business, Stanford University, Hoover Institution, and National Bureau of Economic Research

DANIEL L. RUBINFELD

School of Law and Department of Economics, University of California, Berkeley, and National Bureau of Economic Research

Contents

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)01005-5

Abstract

In this essay, we discuss empirical research on the economic effects of the civil justice system. We discuss research on the effects of three substantive bodies of law—contracts, torts, and property—and research on the effects of the litigation process. We begin with a review of studies of aggregate empirical trends and the important issues involving contracts and torts, both positive and normative. We survey some of the more interesting empirical issues, and we conclude with some suggestions for future work. Because studies involving property law are so divergent, there is no simple description of aggregates that adequately characterizes the subject. In its place, we offer an overview of a number of the most important issues of interest. We describe (selectively) the current state of empirical knowledge, and offer some suggestions for future work. The section on legal process builds on the previous substantive sections. With respect each of the steps, from violation to trial to appeal, we review some of the more important empirical contributions.

Keywords

civil justice, civil litigation, empirical, contract, tort, property, litigation

JEL classification: K11, K12, K13, K41

1. Introduction

In this essay, we discuss empirical research on the civil justice system. Sections 2–4 discuss research on the effects of three substantive bodies of law: contracts, torts, and property. Section 5 discusses research on the effects of the litigation process. Section 5 is closely linked to sections 2–4 because each of the substantive topics (as well as other substantive bodies of law) relies on litigation for enforcement.

Research in all of the areas spans a spectrum from strongly normative to purely descriptive.

Normative economic analyses seek to evaluate the extent to which a law or choice of law apportions costs and benefits to maximize economic efficiency, defined as social surplus (consumer plus producer surplus) less transaction costs. Efficient laws appropriately balance the value of deterrence with the risk-bearing, transactions, and other costs of achieving optimal deterrence. Although it does not fit strictly in an economic framework, normative empirical legal research also examines legal systems' performance on other goals, such as distribution and corrective justice. We discuss this work below as well.

The descriptive arm of the literature is, in general, better-developed than the normative arm. This is largely due to the fact that observational data on the total costs and benefits of law-induced economic or social activity are difficult to obtain. For example, it is impossible to count the total number of contracts that are formed as a result of a body of law—much less the costs and benefits of those contracts to the parties. A typical descriptive study might report trends over time in the outcomes of cases (such as the number of cases, the stage of litigation at which a case is resolved, the probability with which a case is resolved in the plaintiff's or the defendant's favor, and the payment upon resolution of the case); the correlation between case characteristics and case outcomes; or the correlation between measures of the legal environment and case outcomes.

Descriptive analyses, particularly those relating to the effect of a law or system of laws on case outcomes, nonetheless contribute to (normative) policy debates. For example, an assessment of a law's effect on the amount or structure of litigation provides a first step toward an estimate of the magnitude of a law's transactions costs. Descriptive findings about the distributional effects of a law can also be used to assess the law's performance on the overall legal system. Finally, descriptive findings show the extent to which the actual effects of laws conform to their canonical descriptions.

Each of the first two sections of the essay follows a similar format. We begin with a review of studies of aggregate empirical trends and the important issues involving contracts and torts, both positive and normative. We survey some of the more interesting empirical issues, and we conclude with some suggestions for future work. Because studies involving property law are so divergent, there is no simple description of aggregates that adequately characterizes the subject. In its place, we offer an overview of a number of the most important issues of interest. We describe (selectively) the current state of empirical knowledge, and offer some suggestions for future research. The section on legal process builds on the previous substantive sections. With respect each of the steps, from violation to trial to appeal, we review some of the more important empirical contributions.

2. Contract law

2.1. Introduction

Historically, contract disputes were the most common type of civil litigation in the United States. In the mid 19th century, contract cases accounted for the vast majority of all civil litigation (see Galanter, 2001 for an excellent summary of this research). Over the past 150 years, however, contract litigation has become a decreasingly important part of state courts' civil dockets, eclipsed largely by tort cases. Based on case filings in 17 states, the National Center for State Courts reports that by 1995, the number of tort cases exceeded the number of contract cases, although very recent rapid growth in contract cases has reversed this finding (NCSC, 2002).[1]

In a series of reports, the Bureau of Justice Statistics (1996, 2000a) describes the characteristics of contract litigation in greater detail. Contract cases are resolved disproportionately in state courts, reflecting the fact that contract disputes are governed primarily by state law. In 1992, for example, 49,434 contract cases were commenced in federal district courts, as compared to 366,336 cases that were resolved in state courts in the Nation's 75 largest counties. Contract cases reach a conclusion relatively quickly. In 1992, the average time between filing and resolution for a contract case in a studied state court was 13 months, with a median time of 8.3 months; the average time between filing and resolution for a tort case in a studied state court was 19.3 months, with a median time of 13.7 months (Bureau of Justice Statistics, 1995). Relatively few contract cases are between individuals; most cases involved a business or a government institution. In 1992, only 7.9 percent of cases in a studied state court were between individuals, and approximately 60 percent of cases did not involve an individual at all.

We organize our review of empirical research in contract law doctrinally. The first set of studies assesses the effects of contract law versus non-contractual relations on the formation, interpretation, and enforcement of agreements. The second and third sets of studies assess the effects of particular terms of statutory contract law. We group these studies by whether they assess the effects of voluntary terms or mandatory terms (see Ayres and Gertner, 1989, but see the excellent review of empirical contract studies by Korobkin, 2002 for an alternative summary of the literature). Voluntary, or default, rules of contract fill in the gaps in incomplete contracts; parties to a contract are free to contract around them if they wish. Mandatory, or immutable, rules of contract specify the conditions necessary for a contract to be enforceable; they cannot be waived by the parties.

[1] This trend in the mix of litigated cases may not represent the trend in the mix of legal claims, if contract disputes have become disproportionately more likely to be resolved through settlement or arbitration.

2.2. The effects of contract law versus non-contractual relations

The importance of non-contractual relations in business relationships was recognized early by Macaulay (1963). In his study, Macaulay interviewed 68 businesspeople representing 43 companies and six law firms in an open format. He reported strikingly little reversion to formal contract law for the settlement of disputes over contingencies in business agreements. This was not because contingencies were absent from most agreements; his interviews suggested that contingencies were common, but that businesspeople gave little attention ex ante to how they might be resolved. Instead, he found that the value and information inherent in the ongoing relationships that characterized most of the dealings of his sample firms—combined with the high financial and non-financial costs of formal litigation—led firms frequently to resolve disputes through informal means. However, he reported that formal contract law remained important to businesspeople in some important situations: when the terms of agreements were complex; when repeat dealings were unlikely; and when the size or organization of the firm meant that employee who executed the deal needed a formal vehicle to communicate its terms to other members of the firm.

Subsequent work has largely supported Macaulay's view. In two case studies, Bernstein (1992, 1996) examines the use by the diamond and the grain and feed industries of extralegal mechanisms for resolving contract disputes. In her earlier work, she identifies which characteristics of contract law and the diamond industry lead diamond merchants to supplant formal adjudication with their own private arbitration system. She finds that formal contract law is particularly costly to diamond merchants in each of the three efficiency terms discussed above. Uncertainty and delay in recovery increases transaction costs and the costs of the risk of losses; and the use of expectation damages lead to inefficiently too much breach, and therefore too little economic activity ex ante. On the other hand, the small number of traders and powerful reputation effects makes the alternative of relational contracting in this industry particularly low-cost. In her later work, Bernstein examines the private system of adjudication used by the National Grain and Feed Association (NGFA), a trade association of firms and individuals who are active in cash markets for grain and feed. She reports that NGFA arbitrators, as compared to generalist courts, place more weight on the specific terms of parties' agreements than on business norms or customs. She concludes that firms prefer this private system not because non-contractual relations are unimportant to the enforcement of agreements in general, but rather because disputes that require a third party for their resolution are those in which there is little value in the parties' ongoing relationship. Based on a comparable case-study methodology, Lorenz (1999) makes a similar argument: that a drive to enhance efficiency (and therefore gains from trade) leads firms to develop specialized procedural mechanisms to resolve disputes over contracting contingencies.

Using data from a 1995–97 survey of manufacturing firms in Ho Chi Minh City and Hanoi, McMillan and Woodruff (1999a, 1999b) document how firms in a country with poorly developed contract law use extralegal mechanisms. They find that such firms are more likely to offer credit to several types of customers: those with contracts of longer duration, those who are members of "business networks" (through which defaulters can be sanctioned). They also find that third-party intermediaries and ethnic ties are not important mechanisms. Similarly, based on a survey of 36 Bulgarian firms in 1994, Koford and Miller (2006) report that ongoing relationships are important because of the absence of operational contract law in that country.

Johnston (1996) emphasizes how courts have come to respect these non-contractual relationships and the benefits they create. Based on a sample of 25 opinions reported by the UCC reporting service that involved a litigated issue surrounding the statute of frauds—the requirement that contracts involving the sale of goods exceeding $500 must, with certain exceptions, be in writing to be enforceable—he reports two key findings. First, he reports a strong negative correlation of the existence of a prior relationship between the parties and complete absence of any written agreement. Second, he reports a strong positive correlation of a prior relationship and the judge in the case holding that an exception to the statute of frauds applied.

Recent studies use international data to assess the efficiency effects of contract law versus non-contractual relationships. Djankov et al. (2003) measure the impact of features of 109 countries' formal contract law on likely determinants of both economic efficiency—speed of claims resolution and risk—and of fairness in the resolution of two common types of contract disputes: eviction of a tenant for non-payment of rent and collection of a bounced check. They find that increased procedural formalism—as measured by the use of professional judges and lawyers, and the restrictiveness of evidentiary and legal rules—in a country's law reduces both efficiency and fairness. Johnson et al. (2002), based on surveys undertaken in Poland, Romania, Russia, Slovakia, and Ukraine in 1997, conclude that the effect of formal contract law is important to the development of new relationships among firms and to relationships where "lock-in" is inherently low. Similarly, Bigsten et al. (2000) show that access to formal contract law in six African countries enables firms to engage in more complex transactions. Finally, Kranton and Swamy (1999) document an interesting unintended adverse consequence of contract law: reductions in economic efficiency from crowding out of non-contractual relationships. Based on an historical examination of civil courts in colonial India, the authors report that the introduction of formal contract law enhanced competition in credit markets, leading not only to lower interest rates and greater efficiency, but also to unwillingness of lenders to insure farmers against adverse natural events, leading to potentially lower efficiency.

2.3. The effects of voluntary, or default rules of contract law

Voluntary, or default, rules of contract fill in the gaps in incomplete contracts. Theoretically, default rules only matter to the extent that the Coase theorem does not hold. If transaction costs are sufficiently small, and there are no bargaining costs or other market failures, then parties will contract around default rules to establish whatever terms are optimal for their particular situation. If the Coase theorem does not hold, then contract law should set default rules according to what the parties would have wanted or intended, were the transaction costs of negotiating the omitted terms sufficiently small (Ayres and Gertner, 1989). Empirical studies of the impact of default rules generally follow the theoretical path above. The studies begin with a test of the Coase theorem. Then, to the extent that the Coase theorem is rejected, some of the studies investigate the source of the failure of the Coase theorem, and the efficiency or other implications of a particular choice of default rule.

The literature is divided on the importance of default rules. Some work finds that, at least to a first approximation, the Coase theorem holds in bargaining over contract terms. In an experimental study, Schwab (1988) found that changes in the default terms in a hypothetical labor/management negotiation exercise did not significantly prevent the parties from designing the contract in a way that maximized the sum of their rewards. Although the generalizability of this experiment to actual contract negotiations has been questioned (see Korobkin, 1998, discussed below), other work has found that parties "contract around" default terms in real-world situations. Based on a survey of the general counsels of 182 corporations on a broad range of contracting practices, Weintraub (1992) concludes that firms are able to undo default rules that are not jointly optimal. He finds that three default rules of canonical contract law—- failure to enforce promptly repudiated promises, promises to deliver a good when no comparable good is available at the same price, and promises that would, due to unforeseen events, require the liquidation of the promisor—are only selectively followed in practice. Based on telephone interviews with 25 firms involved in the purchase and sale of goods, Keating (2000) investigates the impact of the default rules specified in Section 2-207 of the Uniform Commercial Code, also known as the "battle of forms" provision. (Section 2-207 governs which contract terms will govern an agreement when parties' offer and acceptance contain conflicting, often standard-form, terms.) He finds that firms have adjusted "fairly well" to the default rules imposed by the Section, and that they "don't seem to worry that much" about it.

Other work finds that the Coase theorem does not hold. In an experimental setting similar to Schwab's, Korobkin (1998) found that changes in the default terms in a hypothetical negotiation over shipping services did affect parties' contracts. Because transaction costs in the exercise were low and bargaining power was equally distributed, he hypothesizes that default terms affect contracting behavior because of "status quo bias," due to loss aversion or regret avoidance. McChesney (1999) investigates the effects of the rule of tortious interference—which imposes tort liability on third parties who induce a promisor to breach. In a model without transaction costs, imposing liability for tortious interference would be inefficient, since the expectation damages imposed on a promisor would be sufficient to provide incentives to breach only when it maximized the joint value of the parties relationship. But if transaction costs make renegotiation of agreements prohibitively costly, or lead damages to be systematically below their (economically efficient) expectation levels, then tortious interference may be efficient. Based on an analysis of 134 cases from the early 20th century alleging tortious interference, McChesney finds that the doctrine affects incentives to breach, and that it is more likely to be imposed in those circumstances in which it enhances efficiency. Based on interviews with construction lenders and general contractors, Mann (1996) assesses the consequences of the default rule of first-in-time creditor priority in the context of construction contracts. He compares the responses of interviewees with experience contracting in Missouri, which imposes an unusual contractor-first creditor priority rule, to the response of interviewees with experience contracting under the traditional default rule, and concludes both that the default rule affects contracting outcomes, and that the contractor-first priority rule offers efficiency advantages. Kahan and Klausner (1997) propose a novel reason why default terms might matter: increasing returns to scale, or network effects, in the use of a given default term. They study how event-risk covenants in 101 investment-grade bond issues varied between 1988 and 1993. They find that learning and network effects exist in corporate contracts, and that these externalities in general can lead privately optimal agreements to be economically inefficient from a social perspective.

2.4. The effects of mandatory rules of contract law

In contrast, mandatory rules of contract law cannot be waived by the parties. In the absence of a market failure, such rules necessarily reduce economic efficiency, because they prevent parties from agreeing on terms that they find may jointly beneficial. However, in the presence of market failures (such as asymmetric information), mandatory rules may be efficient; mandatory rules may serve other (distributive) goals as well.

Empirical studies of the effects of mandatory rules fall into two groups. The first group is largely positive in nature and focused on the effects of the doctrine of promissory estoppel. To be enforceable under canonical contract law, the promisor in an agreement had to receive some consideration in exchange for his promise; under this rule, gratuitous promises were not considered valid. The doctrine of promissory estoppel limited the application of this mandatory rule, making gratuitous promises enforceable when the promisee took an action in reliance on the promise.

The empirical literature disagrees about the importance of promissory estoppel. In an analysis of 222 cases alleging promissory estoppel decided between 1981 and 1985, Farber and Matheson (1985) find that requirement of reliance in the doctrine had been interpreted narrowly, thereby reducing the importance of the requirement of consideration. They suggest that modern contract law generally regards any agreement “in furtherance of economic activity” as enforceable. Other studies find that courts are generally unwilling to waive the consideration requirement in contract law. Pham (1994) reports that the doctrine is used infrequently, not as an exception to the consideration requirement but as an all-purpose qualification to a range of mandatory contract rules. In an analysis of 362 cases decided between July 1, 1994 and June 30, 1996, Hillman (1998) reaches a similar conclusion: that the requirement of reliance in the doctrine has been substantially retained, reducing the limitation it imposes on the requirement of consideration (and other) requirements of contract law.

Other studies are more explicitly normative in nature. Although these studies agree that mandatory terms of contract are important, they disagree over their implications.

Rice (1992), for example, examines 628 state supreme and appellate court decisions from 1900 to 1991, and finds that the application of one class of mandatory contract terms—the implied covenant of good faith—promotes neither economic efficiency nor legitimate distributional goals of contract law. In contrast, Ayres (1995) suggests that mandatory disclosure requirements in contracting would enhance efficiency and other legitimate goals. He sent 203 testers to negotiate for the purchase of a new car according to a fixed script, varying only the race and gender of the testers. Ayres found that black and female testers paid higher prices, and concludes that mandatory disclosures by car dealers (for example, of the average price for which each make of car is sold) would lead both to more efficient markets for new cars and to reduced price discrimination on the basis of race and gender.

3. Torts

3.1. Introduction

Over the past 150 years, the number of tort law cases has expanded relative to its closest relative, contract cases. In a series of reports analogous to those on contract cases, the U.S. Department of Justice, Bureau of Justice Statistics (1995, 1997, 1999, 2000b) describes the characteristics of tort litigation in detail. Like contract disputes, tort cases are generally governed by state law and resolved in state courts. In 1994, for example, 41,166 tort cases were resolved in federal district courts, compared to 378,314 cases resolved in state courts in the Nation's 75 largest counties in 1992. Tort cases take significantly longer to resolve than contract cases. In 1992, the average time between filing and resolution for a tort case in a studied state court was 19.3 months, with a median time of 13.7 months (Bureau of Justice Statistics, 1995). Because tort litigation primarily involves damages related to a personal injury, the vast majority of tort cases (unlike contract cases) report an individual as the plaintiff (94 percent). Tort cases disproportionately involve auto accidents. In 1992, 60.1 percent of tort cases in the Nation's 75 largest counties involved an auto accident; the second, third, and fourth most common types of tort cases were premises liability (17.3 percent), medical malpractice (4.9 percent), and products liability (3.4 percent) (Bureau of Justice Statistics, 1995).

There is disagreement over the causes of the long-run growth in tort cases. One view, elaborated by Schwartz (1981, 1992) and Priest (1988) among others, emphasizes that the expansion of the scope of tort law is due substantially to the rise of judicial and public-policy activism in the 1960s and 1970s. According to this view, the shifting of the costs of accidental injury from (largely individual) plaintiffs to (largely institutional) defendants has occurred because of increasing acceptance of two positions: (1) institutional defendants are generally more able to prevent accidents, (2) it is good policy to provide insurance through the tort system for the costs of those accidents that can not be prevented.[2] Another view, elaborated by Galanter (1996) and Eaton et al. (2000) among others, is that the growth in tort claims has been slower, more gradual, and only weakly related to reforms of the 1960s and 1970s, after appropriate adjustment for population and other factors.

Most empirical research in tort law, however, is concerned with assessment of the effects of tort law on distributional goals and economic efficiency, where efficiency (as in contract law) involves the tradeoff between transactions costs, incentives for appropriate activity, and least-cost allocation of the risk of harm or loss. We organize our review of empirical studies of tort law by type of accident, because the most important empirical issues differ for different types of accidents. Subsection 4.2 reviews the effects of tort in medical malpractice cases. The important empirical issue in this context is assessing the extent to which the cost and asymmetries of information in markets for medical care interact with conventional tort doctrine to create incentives for "defensive medicine"—the use of treatments with minimal medical benefit, or the refusal to use treatments with substantial benefit, out of fear of legal liability. Subsection 4.3 reviews the literature on the effects of tort law in auto accident cases. The most important empirical issue in this context is comparing the transaction costs, incentives for accident avoidance, and compensation paid under the traditional tort system versus a no-fault system administered through mandatory first-party insurance. Subsection 4.4 reviews work on product liability. Subsections 4.5 and 4.6 conclude with discussion of two important policy issues that span across sectors: the issue of punitive damages and the issue of mass torts. We leave the discussion of the role of the tort system in the environmental and labor law to other chapters; see Dewees, Duff, and Trebilcock (1996) for an excellent review of the empirical and theoretical issues in these contexts, and for an alternative review of the issues covered in this section. We also omit discussion of constitutional tort litigation, a federal cause of action for civil damages and injunctions against state officials who deprive others of constitutional and certain federal statutory rights; see Eisenberg and Schwab (1987) for an empirical study of this type of claim.

3.2. Medical malpractice[3]

Injuries due to malpractice create substantial economic costs. According to the Institute of Medicine (2000), medical errors are the leading cause of accidental death in the United States: estimates of the number of deaths due to medical errors in 1997 range from 44,000 to 98,000. Medication errors alone account for approximately 7,000 deaths per year, exceeding the number of deaths due to workplace injuries. Weiler et al. (1993) find that nearly one percent of hospital admissions in New York State in 1984 involved an injury due to negligent care; the proportion of serious injuries due to negligence was even higher.

[2] But see Shavell (1987) for an analysis that shows the limited circumstances under which this hypothesis is likely to be true.

[3] Pieces of this section have appeared previously in Kessler and McClellan (1996).

In theory, the liability system should provide physicians with the incentives to take the socially optimal amount of care against medical injury. By making doctors responsible for the costs of injuries that they negligently cause, the negligence rule should lead doctors' private decisions as to whether and how to practice medicine to reflect society's overall interests by leading them to balance in expectation terms the social benefits of medical care, the costs of precaution, and the costs of negligence.

In practice, several factors—most notably, the pervasiveness of insurance in the health care sector—drives a wedge between these seemingly-sensible incentives and socially optimal medical decision-making. On one hand, market failures may create incentives for too little precaution. First, because of transaction costs or asymmetries of information, patients often do not file malpractice claims even when there is evidence of a negligent medical injury (Harvard Medical Practice Study, 1990). Second, because malpractice liability insurance is at most weakly experience-rated (Sloan, 1990), physicians may bear little of the costs of patient injuries from malpractice. (However, physicians bear significant uninsured, nonfinancial costs of patient injuries, including the value of lost time and emotional energy in responding to a malpractice claims, Office of Technology Assessment, 1993).

On the other hand, market failures can create incentives for too much precaution, or "defensive medicine." The practice of "defensive medicine" can take two forms: positive defensive medicine and negative defensive medicine. Positive defensive medicine involves the supply of care that is relatively unproductive for patients; negative defensive medicine involves declining to supply care that is relatively productive for patients.

Positive defensive medicine is driven by moral hazard from health insurance, which means that neither patients nor physicians bear the most of the costs of care in any particular case. The costs of precautionary services financed through health insurance are generally larger than the uninsured cost of the physician's own effort. Doctors and patients make decisions that balance the costs of precaution that they bear against the costs imposed on them by the malpractice system. Thus, even if medical malpractice tort law allocated the burden of medical injuries with neither errors nor transaction costs, insensitivity to the true costs of care would lead physicians and their patients to prefer to take socially excessive precautions against injuries. The added burden due to errors and transaction costs only intensifies this effect. As Danzon (1991) observes, the fact health care providers' precautionary behavior may be ex post difficult to verify may give them the incentive to take inefficiently too much care. Craswell and Calfee (1986) prove that such excessive care results from the all-or-nothing nature of the liability decision: small increases in precaution above the optimal level can result in large decreases in expected liability.

Negative defensive medicine is driven by the fact that patients reap substantial surplus from medical care for which they can not compensate providers, while providers bear malpractice risk for which they can not charge patients. If doctors weigh the malpractice downside of a course of care against only a fraction of the upside, then they may withhold treatments that may be in patients' best interests.

Both forms of defensive medicine may be much more economically important than the costs of awards and settlements imposed by the malpractice system. If the costs of precaution borne by patients and physicians account for a small share of the total costs, but the costs imposed by the malpractice system are roughly proportional to the costs of awards and settlement payments, then a given change in malpractice awards would induce a much greater change in the costs of medical care. Thus, although the direct cost of malpractice tort awards, including transaction costs, is relatively insignificant at roughly one percent of medical spending (Danzon, 2000), the indirect costs of the malpractice tort system, in terms of wasteful medical treatment or loss of life due to medical errors, is likely to be far greater.

We organize our review of empirical assessment of the effects of malpractice tort law in three parts (but see Danzon, 2000 for an excellent and comprehensive alternative review). The first arm of the literature surveys physicians about their opinion of the role of the malpractice system in determining medical treatments (Reynolds et al., 1987; Moser and Musaccio, 1991; Office of Technology Assessment, 1994; Klingman et al., 1996). Although surveys indicate that physicians believe that the existing malpractice system leads to economically inefficient medical care decisions, surveys only provide information about physicians' self-reported perceptions of the effects of law: they do not measure behavior in real situations.

We discuss the remainder of the literature in two subsections below. The second arm of the literature estimates the effects of the characteristics of claims and of malpractice laws on compensation and other dimensions of "malpractice pressure"—the incentives for hospitals and physicians to shield themselves against the costs of legal liability. Although estimates from this literature have contributed to our understanding about how malpractice law works in practice, they only show how the malpractice system affects incentives, not how it affects treatment behavior—and therefore the cost and quality of care. The third arm of the literature seeks to directly measure the effects of variation in malpractice law and malpractice pressure across states on actual medical care decisions, health care costs, and patient health outcomes. This literature finds that, at least in areas with high levels of malpractice pressure, reductions in liability lead to increases in economic efficiency—that the tort system on net creates incentives for defensive medicine—although studies differ in their assessment of the magnitude of this effect. Because of the lack of empirical research, we omit discussion of two important alternatives to the tort system for apportioning damages from medical injury: no-fault insurance (but see Studdert et al., 1997; Studdert, Brennan, and Thomas, 2000; and Studdert and Brennan, 2001 for a simulation of the costs and benefits of a hypothetical no-fault system) and contract (which, because of the pre-existing relationship between the parties, is a viable alternative to tort (e.g., Shavell, 1987); see Havighurst, 1995 for a theoretical evaluation of apportionment of medical liability through contract).

3.2.1. Effects of malpractice law on malpractice pressure

Simple statistical indicators suggest that the role of the malpractice system in health care has grown over the past 40 years. As Danzon (2000) points out, the number of malpractice claims per physician and the award paid per claim increased rapidly in the US from the 1960s to the 1980s. Claim frequency increased at more than 10 percent per year, reaching a peak of 17 claims per 100 physicians in 1986. Awards paid per claim increased at roughly twice the rate of inflation (Danzon, 1986), with some evidence of even greater growth for the most costly cases (Shanley and Peterson, 1987). In turn, malpractice liability insurance markets experienced two "crises," one in the mid-1970s and one in the mid-1980s, in which prices for malpractice insurance skyrocketed and availability of insurance contracted (see Danzon, 2000, section 4.2, for a cataloguing of the literature on malpractice insurance). Taken together, these factors led states' legislatures to change their laws governing medical malpractice claims to reduce liability.

Numerous studies have investigated the effects of the characteristics of malpractice claims and tort law on measures of malpractice pressure such as malpractice claims rates, the awards paid to and other measures of the disposition of claims, and malpractice liability insurance premiums. This literature reports three main findings. First, economic loss, rather than fault, is consistently the most important characteristic of claims in determining the probability and size of award (Danzon and Lillard, 1983; Farber and White, 1991; Brennan et al., 1996). Second, the changes in state laws designed to reduce liability significantly reduced the incentives to supply precautionary care. Danzon (1982, 1986) and Sloan et al. (1989) find that tort reforms that cap physicians' liability at some maximum level or require awards in malpractice cases to be offset by the amount of compensation received by patients from collateral sources[4] reduce payments per claim. For example, based on 1975–1978 data, Danzon (1982) reports that states enacting caps on damages had 19 percent lower awards, and states enacting mandatory collateral source offsets had 50 percent lower awards. Based on 1975–1984 data, Danzon (1986) reports that states enacting caps had 23 percent lower awards, and states enacting collateral source offsets had 11 to 18 percent lower awards. Based on 1975–1978 and 1984 data, Sloan et al. (1989) find that caps reduced awards by 38 to 39 percent, and collateral source offsets reduced awards by 21 percent. Danzon (1986)

[4] Reforms requiring collateral-source offset revoke the common-law default rule which states that the defendant must bear the full cost of the injury suffered by the plaintiff, even if the plaintiff were compensated for all or part of the cost by an independent or "collateral" source. Under the common-law default rule, defendants liable for medical malpractice always bear the cost of treating a patient for medical injuries resulting from the malpractice, even if the treatment were financed by the patient's own health insurance. Either the plaintiff enjoys double recovery (the plaintiff recovers from the defendant and his own health insurance for medical expenses attributable to the injury) or the defendant reimburses the plaintiff's (subrogee) health insurer, depending on the plaintiff's insurance contract and state or federal law. However, some states have enacted reforms that specify that total damages payable in a malpractice tort are to be reduced by all or part of the value of collateral source payments.

also finds that collateral-source-rule reforms and statute-of-limitations reductions reduce claim frequency. Based on data from malpractice insurance markets, Zuckerman et al. (1990) and Barker (1992) reach similar conclusions: Zuckerman et al. find that caps on damages and statute-of-limitations reductions reduce malpractice premiums, and Barker finds that caps on damages increase profitability. Third, the two reforms most commonly found to reduce payments to and the frequency of claims, caps on damages and collateral source rule reforms, share a common property: they *directly* reduce expected malpractice awards. Caps on damages truncate the distribution of awards; mandatory collateral source offsets shift down its mean. Other malpractice reforms that only affect malpractice awards *indirectly*, such as reforms imposing mandatory periodic payments (which require damages in certain cases to be disbursed in the form of annuity that pays out over time), have had a less discernable impact on liability and hence on malpractice pressure.

Taken alone, estimates of the impact of reforms on frequency and severity from these analyses are only the first step toward answering the policy question of interest. They provide evidence only of the effects of legal reforms on doctors' incentives; they do not provide evidence of the effects of legal reforms on doctors' behavior. Identifying the economic efficiency of precautionary behavior due to legal liability requires a comparison of the response of costs of precaution and the response of losses from adverse events to changes in the legal environment.

3.2.2. Effect of malpractice law and pressure on the components of economic efficiency

The third arm of the literature investigates how treatment decisions and patient health outcomes respond to malpractice pressure. Early work estimates the effect of physicians' actual exposure to malpractice claims to clinical practices and outcomes (Rock, 1988; Harvard Medical Practice Study, 1990; Localio et al., 1993; Baldwin et al., 1995). Rock, Localio et al., and the Harvard Medical Practice Study find results consistent with defensive medicine; Baldwin et al. do not. However, concerns about unobserved heterogeneity across providers and across small geographic areas qualify the results of all of these studies. These studies use frequency of claims or magnitude of insurance premiums at the level of individual doctors, hospitals, or areas within a single state over a limited time period to measure malpractice pressure. Because malpractice laws within a state at a given time are constant, the measures of malpractice pressure used in these studies arose not from laws but from primarily unobserved factors at the level of individual providers or small areas, creating a potentially serious problem of selection bias. For example, the claims frequency or insurance premiums of a particular provider or area may be relatively high because the provider is relatively low quality, because the patients are particularly sick (and hence prone to adverse outcomes), because the patients had more "taste" for medical interventions (and hence more likely to disagree with their provider about management decisions), or because of many other factors; the sources of the variation in malpractice pressure are unclear and probably multifactorial. All of these factors are extremely difficult to capture fully in observational datasets, and could lead to an apparent but non-causal association between measured malpractice pressure and treatment decisions or outcomes.

More recent work seeks to address this endogeneity problem by identifying the effects of malpractice pressure with variation across states and over time in malpractice law reforms, which are arguably exogenous (but see Danzon, 2000 for a critique of this assumption). Much of this work has investigated the consequences of malpractice pressure for positive defensive medicine. Kessler and McClellan (1996) use longitudinal data on essentially all elderly Medicare beneficiaries hospitalized with serious cardiac illness from 1984, 1987, and 1990, matched with information on the existence of direct and indirect law reforms from the state in which the patient was treated. They found that reforms that directly limit liability—such as caps on damages—reduced hospital expenditures by 5 to 9 percent in the late 1980s, with effects that are greater for ischemic heart disease (IHD) than for heart attack (AMI) patients.[5] In contrast, reforms that limit liability only indirectly were not associated with any substantial expenditure effects. Neither type of reforms led to any consequential differences in mortality or the occurrence of serious complications. The estimated expenditure/benefit ratio associated with liability-pressure-induced intensive treatment was over $500,000 per additional one-year survivor, with comparable ratios for recurrent AMIs and heart failure. Thus, treatment of elderly patients with heart disease does involve defensive medical practices, and limited reductions in liability can reduce this costly behavior.[6]

Two recent studies identify the mechanism through which direct reforms affect physician behavior, in order to help predict whether existing reforms under new market conditions, or new and untried types of reforms, will have similar effects. Kessler and McClellan (2002b) match longitudinal Medicare data with law reforms and data on health insurance markets to explore the ways in which managed care and liability reform interact to affect treatment intensity and health outcomes. They report that direct reforms reduce defensive practices in areas both with low and with high levels of managed care enrollment. Managed care and direct reforms do not have long-run interaction effects that are harmful to patient health. However, at least for patients with less severe cardiac illness, managed care and direct reforms are substitutes, so the reduction in defensive practices that can be achieved with direct reforms is smaller in areas with high managed care enrollment.

Kessler and McClellan (2002a) integrate four unique data sources to illuminate how reforms affect malpractice pressure, and how reform-induced changes in the incentives provided by the liability system affect treatment decisions, medical costs, and health outcomes. That paper matches by state and year the longitudinal Medicare data discussed above (updated to include all years from 1984–1994) with data on law reforms;

[5] Because IHD is a less severe form of illness, IHD patients may have more “marginal” indications for intensive treatment, leading to a greater scope for defensive practices.

[6] Dubay et al. (1999) confirm that defensive practices exist in non-elderly populations, although they report that the costs of defensive medicine in obstetrics are small.

physician-level data on the frequency of malpractice claims from the American Medical Association's Socioeconomic Monitoring System (AMA SMS); and malpractice-claim-level data from the Physician Insurers Association of America on claim costs and claim outcomes. They report that although direct reforms improve medical productivity primarily by reducing malpractice claims rates and compensation conditional on a claim, other policies that reduce the time spent and the amount of conflict involved in defending against a claim can also reduce defensive practices substantially. For example, at least for elderly heart disease patients, an untried reform that reduced the legal-defense burden on physicians and hospitals by one-quarter—which is within the range of policy possibilities—could be expected to reduce medical treatment intensity by approximately 6 percent, but not to increase the incidence of adverse health outcomes. In the same population, a policy that expedited claim resolution by six months across-the-board could be expected to reduce hospital treatment costs by 2.8 percent, without greater adverse outcomes. This finding is consistent with Kessler and McClellan (1997), which reports broad differences in physicians' perceptions of the impact of malpractice pressure in states with and without liability reforms.

Other work has investigated the consequences of malpractice pressure for negative defensive medicine. For example, Dubay et al. (2001) find that a decrease in malpractice premiums that would result from a feasible policy reform would lead to a decrease in the incidence of late prenatal care by between 3.0 and 5.9 percent for black women, and between 2.2 and 4.7 percent for white women. However, although they found evidence that malpractice pressure was associated with greater delay and fewer prenatal visits, they found no evidence that this negatively affected infant health. More recent work finds substantial evidence of negative defensive medicine. Hellinger and Encinosa (2003) report that the supply of physicians was approximately 12 percent greater in states with caps on non-economic damages, as compared to states without them.

3.3. Auto accidents

After medical errors, auto accidents are the second-leading cause of accidental death in the United States, with just over 40,000 deaths per year (U.S. Department of Transportation, 2003). Adding the costs of less severe injuries increases the estimated magnitude of costs of auto accidents. Roughly two-thirds of all civil litigation involves an automobile accident (Rolph et al., 1985). According to the Urban Institute (1991), auto accidents cause a total of $420 billion in total damage every year. Approximately one-quarter of this amount, or $100 billion, is passed through the automobile liability insurance system.

Empirical research in the auto accident context has focused on comparisons of the performance of tort versus no-fault systems (but see Vickrey, 1968; Wittman, 1986; White, 1989; Kessler, 1995; and Edlin, 2003 for comparisons of the incentives for precaution provided by different systems of tort law). Under auto no-fault systems, some proportion of losses from accidents are compensated by a driver's own (mandatory) insurance regardless of negligence, with the remainder apportioned by tort law. Under tort law, losses from accidents are borne by the party at fault. The empirical question generally takes the following form: can no-fault systems achieve the same level of deterrence as tort systems, but with lower transactions costs and significantly better performance on generally-accepted compensation goals?

There is substantial agreement that no-fault systems administer compensation for auto accidents with much lower transaction costs than does the tort system. According to Carroll et al. (1991), costs of litigation, settlement, and other administration of compensation in a typical tort liability system amounted to about 33 percent of the cost of injuries covered by insurance; this excludes the publicly-financed costs of administering the civil justice system, which are substantial (Kakalik and Robyn, 1982). Under a typical no-fault plan, they find that transaction costs would be reduced by about 39 percent. On a base of $100 billion per year (Carroll et al.), this amounts to a lower-bound estimate of efficiency improvements of $13 billion per year (100*.33*.39).

Carroll et al. (1991) also show that no-fault would deliver compensation faster and make it more closely track economic loss. Under a typical tort system, claimants receive initial compensation payments 181 days after the accident; under a typical no-fault system, claimants receive initial compensation payments 116 days after the accident, or 36 percent faster. In addition, under a typical tort system, claimants with less serious injuries tend to receive more than their economic loss (see also Carroll and Abrahamse, 1999, and the work cited therein, for evidence on the extent of overcompensation due to fraud and abuse), with fully 62 percent of all injured people receiving more than their economic loss; claimants with more serious injuries tend to receive less than their economic loss, with 27 percent of injured people receiving less than their economic loss. By comparison, under a typical no-fault system, only 22 percent of injured people receive more than their economic loss, and 16 percent receive less than their economic loss.

From a theoretical perspective, the overall impact of mandatory no-fault on accident rates is ambiguous. In theory, mandatory no-fault compensation for damages unambiguously reduces the incentives to take care, because drivers do not directly bear the costs of their negligence. In practice, however, several factors might mitigate this result. Under a typical no-fault system, a driver would still have incentives to take care to avoid injury, to avoid criminal and traffic penalties, and to avoid the premium increases that would likely accompany an accident under a conventionally-experience-rated no-fault insurance policy (Loughran, 2001).

The empirical papers investigating the effects of no-fault on accidents share a common research design. They use data at the state/year level on the rate of fatal auto accidents; the characteristics of states and their populations, such as weather, alcohol consumption, real income, miles of roads, the mean and variance of vehicle speeds, the proportion of vehicle registrations that are trucks, hospitals per square mile, and the age, education, and employment statuses of their population; and state auto accident law, including whether or not the state had mandatory no-fault insurance, the scope of losses covered by the no-fault system; and other laws governing the apportionment of damages from auto accidents. They estimate the effect of no-fault on the fatal accident rate by comparing the trend in the fatal accident rate in (all or a subset of) the seventeen states that adopted no-fault insurance between 1971 and 1990 to the accident rate in all other states, holding constant other covariates.

These papers generally find that no-fault leads to greater numbers of fatal accidents, with some papers based on earlier data finding no effect. Based on 1967 to 1975 data, Landes (1982) finds that no-fault systems with relatively low "tort thresholds" lead to increases in fatal accident rates of approximately 4 percent, and (more expansive) no-fault systems with high tort thresholds lead to increases in accident rates in excess of 10 percent. In contrast, follow-up work by Kochanowski and Young (1985) using data from 1975, 1976, and 1977, and by Zador and Lund (1986), using data from 1976 to 1980, found no effect of no-fault on accident rates.

Later work, and work based on data from other countries, has consistently found that the decreased incentives created by no-fault increases the accident rate. Cummins et al. (2001) use data from 1968 to 1994. Using simple ordinary least squares regression ("OLS")models, they find that no-fault increases the accident rate statistically significantly by 5 percent; in simultaneous-equations models of the joint process by which states adopt no-fault and the effect of no-fault, they find that no-fault increases the accident rate by as much as three times this amount. Cohen and Dehejia (2003) use data from 1970 to 1998 and find in OLS models that no-fault leads to an increase in traffic fatalities of approximately 6 percent. Sloan et al. (1995) report that tort liability accomplishes this reduction in fatal accidents by reducing the prevalence of one of the key determinants of auto accidents—binge drinking.[7] McEwin (1989), using state data from Australia and New Zealand from 1970 to 1981, and Devlin (1990), using provincial data from Canada from 1967 to 1984, also find positive effects of no-fault on the fatal accident rate of 16 and 9 percent, respectively.

In a series of papers, O'Connell and various coauthors (O'Connell and Joost, 1986; O'Connell et al., 1993, 1995, 1996), Abrahamse and Carroll (1995), and Carroll and Abrahamse (1999) argue that a "choice" system of auto insurance offers a viable compromise that addresses the limitations of both mandatory no-fault and tort. Under a choice system, drivers choose whether to be insured under a traditional tort law or under an administrative no-fault system. Those who choose the tort system retain traditional tort rights and responsibilities, but those who choose no-fault can neither recover for nor are liable for those losses that are designated to the no-fault system. However, as Kabler (1999) points out, none of this work explains why a choice system would avoid the adverse accident effects of no-fault systems. And, adverse selection by (high-premium) risky drivers into the no-fault system may decrease the value of retaining tort coverage so much that a choice system would collapse into mandatory no-fault.

[7] See Chaloupka et al. (1993), Sloan et al. (1994), and Ruhm (1996) for a comparison of the effectiveness of tort liability versus other approaches for deterring fatalities due to drinking and driving.

3.4. Products liability

Losses from injuries from products are controlled differently than are losses from other types of accidents. First, regulatory mechanisms other than tort provide important incentives for product safety. The Consumer Products Safety Commission, The Occupational Health and Safety Administration, The Food and Drug Administration, The Federal Aviation Administration, and the Environmental Protection Agency all regulate the safety of products. Second, over the past three decades, most US jurisdictions have adopted strict liability for injuries from products (Keeton et al., 1984). This means that manufacturers and retailers are liable for damages caused by product defects, even if they exercised all possible care in the preparation and sale of the product (Restatement (Second) of Torts, Section 402A (1977)).

This combined regulatory/liability system could provide incentives for either too much or too little precaution from the perspective of economic efficiency. On one hand, the joint regulatory/liability system may provide incentives for economically efficient precautionary care decisions as well as or better than the basic “negligence rule.” For example, if regulatory agencies’ monitoring is infrequent and ability to impose sanctions is limited, and transaction costs limit the tort system’s ability to provide sufficient incentives for safety, then a combined regulatory/liability system may be socially preferred to either system standing alone. On the other hand, the combined costs imposed by a system of strict liability for product-related accidents and the regulatory system may lead to too much precaution. This excessive precaution can take two forms: socially insufficient or misdirected product innovation and insufficient or misdirected product use (e.g., activity levels).

We organize our review of the empirical literature in the same three parts as in medical malpractice (see also Litan, 1991 and Congressional Budget Office, 2003 for excellent alternative reviews). The first arm of the literature seeks to assess the impact of law and regulation using surveys and case studies. Some of this work finds product liability law has small or ambiguous effects on innovation and activity level (e.g., Eads and Reuter, 1985; Johnson, 1991); some finds evidence of adverse efficiency effects (e.g., Mackay, 1991); and some finds evidence of both socially constructive and adverse effects (e.g., Garber, 1993). However, because of the context-specific nature of this work, the findings are often difficult to generalize beyond the firms or individuals featured in the study.

We discuss the remainder of the literature in two subsections below. As in medical malpractice, we first describe work that estimates the effects of the characteristics of claims and of products liability laws on compensation and other dimensions of “liability pressure”—the incentives for product safety. Then, we discuss work that seeks to directly measure the effects of variation in products liability law and liability pressure on the three components of the efficiency costs and benefits of products liability law: accidents and deaths; the costs and availability of products; and measures of innovation, research, development, and productivity.

3.4.1. Effects of products liability law on liability pressure

Numerous studies have sought to assess whether liability pressure in the products area has been rising or falling nationally over the past 20 years. Indices constructed by Eisenberg and co-authors from filed cases conclude that liability pressure has been falling (Henderson and Eisenberg, 1990; Eisenberg and Henderson, 1992; Eisenberg et al., 1996; Eisenberg, 1999). However, these analyses do not account for liability pressure created by claims that were not litigated, i.e., that did not result in a court filing. To the extent that the selection of cases has been changing over time, indices based only on filed cases will be biased (but see Eisenberg and Henderson, 1992 for an argument why selection bias is unimportant to their findings). Indeed, indices based on insurance premiums and other data that include all payments to claimants, whether or not they result in a court case, find a dramatic increase in the cost of products liability (e.g., Viscusi, 1991). The expansion of compensation through the tort system reported in these studies is consistent with the findings of Henderson (1991), who finds "fairness" more important than "efficiency" in judges' reasoning in products liability cases.

Although this work provides background about trends over time in the incentives provided by the tort system, it does not identify how products liability law affects incentives for care. Since many determinants of liability pressure other than law have changed over time, national trends reflect some combination of these effects and changes in law. Three studies provide evidence that limitations on liability reduce liability pressure. Viscusi (1990) uses data from products liability insurance markets from 1980 to 1984. He finds that state statutes designed to reduce liability pressure—including those that provide a state-of-the-art defense, a statute of limitations for producer liability, and collateral-source offsets and other damages restrictions—reduce products liability premiums, reduce insurer loss ratios, and increase insurance availability. Manning (1994) estimates the effect of liability costs on product prices by comparing the difference in trends from 1950–1989 in the prices of the DPT vaccine (which confers immunity from diphtheria, pertussis (whooping cough), and tetanus) and the DT vaccine (which confers immunity from diphtheria and tetanus only). He argues that this identifies the impact of liability costs because the DPT and DT vaccines have roughly comparable costs of production, but the DPT vaccine has significantly higher liability costs (because of the possibility of rare but serious complications from its pertussis component). He finds that the wholesale price of the DPT vaccine rose by over 2,000 percent during the study period, and that 96 percent of the price increase was due to litigation costs. Manning (1997) estimates the effect of liability costs on the relative U.S./Canadian prices charged by the manufacturer of 121 of the 200 most prescribed prescription drugs in 1990. He compares the relative prices of drugs with substantial liability risks, as measured by litigation history, controlled substance designation, relationship to vaccine liability costs, risk-assessment survey response, and drug-reference book risk rating, to the relative prices of drugs without such risks, holding other determinants of relative prices constant. He finds that liability risk roughly doubles the average US/Canadian price differential, and increases the median price differential by about one-third.

3.4.2. Effects of products liability law and liability pressure on economic efficiency

The third arm of the literature investigates how products liability law affects economic efficiency. Some studies identify the efficiency effects of products liability law by comparing national trends in accident rates after versus before periods of expansion of product liability pressure. Priest (1988) compares trends in the rate of accidental deaths and injuries from products for the late 1970s to trends for the 1960s and early 1970s. He finds an increase in product deaths and injuries in the later period, and concludes that the expansion of product liability had minimal (if any) beneficial effects on accident avoidance. Martin (1991) analyzes data from the aviation industry using a similar methodology. He finds sharper declines in the aviation accident rate from 1950–69 than from 1970–89.

Like the work on trends over time in the incentives provided by the tort system, though, this work is only suggestive, since many other determinants of the accident rate (such as regulatory policy) may have been changing contemporaneously with aggregate trends in products liability pressure. Two studies attempt to control for such factors. Campbell et al. (1998) use data on labor productivity (gross state product per worker) by industry from 1970 to 1990. They compare trends in productivity in states that adopted liability law reforms, such as caps on damage awards, to trends in productivity in states that did not, holding constant fixed differences across states and the time-varying political and economic characteristics of states. They find that states that changed their liability laws to reduce levels of liability experienced greater increases in productivity that states that did not. Viscusi and Moore (1993) use data from 1980–84 from the Profit Impact of Marketing Strategies (PIMS) study, containing balance sheet and income statement items for distinct lines of business from large firms, matched by 3-digit SIC industry code with data from the same period on product liability cost measures from the Insurance Services Office. They report a positive effect of liability on product innovation at low to moderate levels of liability pressure, but a negative effect of liability on product innovation at high levels of liability. In addition, they find that liability pressure has a greater effect on product innovation than on process innovation.

3.5. Punitive damages

The economic effects of punitive damages—those in excess of what is necessary to compensate a plaintiff for his economic and noneconomic losses—is an important policy issue that applies to each of the four types of tort cases discussed above. On one hand, punitive damages may be necessary to induce optimal precautionary behavior in two situations: those in which injurers' behavior is undetected, and those in which can escape liability, and those in which the benefits enjoyed by injurers should be excluded from social welfare (Polinsky and Shavell, 1998). On the other hand, if punitive damages are awarded in other situations, or in amounts disproportionate to the probability of defendants' escaping liability, then they may lead to socially excessive precaution.

We organize our review of the empirical literature in the same three parts as in medical malpractice and products liability (see also Robbennolt, 2002 for an excellent alternative review of the empirical literature on punitive damages). The first arm of the literature seeks to assess the impact of punitive damages using surveys. Baker (1998) reports that lawyers believe that statutory restrictions on punitive damages would be at best only partially effective, because juries will inflate compensatory damages to offset their reduced ability to impose punitive damages. Launie et al. (1994) report that a sample of 100 small business owners and 100 corporate counsel believe that punitive damages increase estimated settlement payments by 13 percent. However, as in the case of surveys in products liability, the context-specific nature of this work makes findings difficult to generalize beyond the firms or individuals featured in the study.

We discuss the remainder of the literature in two subsections below. As in medical malpractice, we first discuss work that estimates the effects of punitive damages on measures of "liability pressure"—the incentives for safety. Then, we discuss work that seeks to directly measure the effects of punitive damages on the efficiency of tort law.

3.5.1. *Effects of punitive damages on liability pressure*

Numerous studies have sought to assess whether punitive damages have an important effect on the incentives for accident avoidance. Most of this work has investigated the frequency and magnitude of punitive damages awards in litigated cases (Peterson et al., 1987; GAO, 1989; Daniels and Martin, 1990; Hensler and Moller, 1995; Moller et al., 1997; Eisenberg et al., 1997, 2002). The studies report similar overall probabilities of a punitive damage award, ranging from 4.7 percent of jury verdicts in Cook County, Illinois and California from 1980–84 (Peterson et al., 1987), to 4.9 percent of 25,627 jury verdicts from state trial courts in eleven states from 1981–85 (Daniels and Martin, 1990), to 6 percent of 6,053 trials in the Civil Trial Court Network sample of state trial courts in 45 counties in 1991–92 (Eisenberg et al., 1997). However, punitive damages are awarded substantially more frequently in certain types of cases, including those involving financial injuries (Moller et al., 1997); business/contract cases (Peterson et al., 1987); and fraud and intentional tort cases (Peterson et al., 1987; Eisenberg et al., 2002). The median punitive damage award is roughly equal to the median compensatory damage award, but the mean punitive damage award is much larger (Eisenberg et al., 1997), because of a small number of very large punitive damage cases. The two RAND studies that examine punitive damage awards over time (Peterson et al., 1987; Hensler and Moller, 1995) find that although the frequency of punitive damage awards in trials have remained roughly constant, total punitive-damage dollars awarded has risen substantially. However, as in the products liability context, the frequency, severity, or trend in frequency or severity of punitive damages awarded in litigated cases does not account for liability pressure created by claims that were not litigated, i.e., that did not result in a court filing. As Polinsky (1997) explains, because litigated cases are such a small and characteristically unrepresentative sample of all claims, extrapolating the results from these studies to the universe of claims is not possible without substantial additional assumptions that are unlikely to be correct.

Other studies have sought to identify the liability pressure created by punitive damage awards with broader measures. Koenig (1998) reports on several studies conducted by the Texas Department of Insurance (TDI) analyzing all closed liability insurance claims in that state. Insurance adjusters review each claim in the TDI data and apportion settlement payments into compensatory and punitive damages. In these studies, the TDI estimates the effect that statutory restrictions on punitive damages adopted by Texas in 1995, requiring plaintiffs to show clear and convincing evidence of malice to obtain punitive damages. TDI concludes that punitive damages accounted for between 10 and 16 percent of all claims payments, depending on the size of the claim and the time period and that the restrictions on punitive damages would reduce liability insurance premiums by 4.5 percent, or approximately $428 million.

As with litigated cases, the liability pressure created by punitive damages varies substantially across case types. Using 1983 data on commercial bodily injury (other than medical malpractice) insurance claims of $25,000 or more, the Insurance Services Office (1988) shows that less than one percent of payments to such claims are attributable to punitive damages. Karpoff and Lott (1999) use event studies to estimate the impact on firms' stock prices of several key Supreme Court decisions, congressional actions, and specific lawsuits that changed firms' exposure to punitive damages liabilities. They find little evidence that the announcement of changes in the legal environment affect valuations, but substantial evidence that the announcement of specific lawsuits affect valuations by more than the direct cost of the punitive damage award.

A new literature based on jury experiments investigates whether punitive damages are awarded in situations and in amounts that would create the incentives for optimal precaution (Baron and Ritov, 1993; Sunstein et al., 1998; Viscusi, 2001a, 2001b, 2002). The experiments ask jury-eligible participants to report the punitive damages that they would award in hypothetical fact situations under different sets of jury instructions or statutes. The experiments also ask participants about the reasoning that they used to reach their decision. The studies find that punitive damages depend neither on the defendant's ability to escape detection (as would be necessary to achieve economic efficiency) nor on other case characteristics that would be necessary to achieve most notions of distributional "fairness". In addition, the studies conclude that offering juries additional guidance by increasing the specificity of jury instructions (e.g., by providing them with a mathematical formula for determining punitive damages) have little effect (see especially Viscusi, 2001b). Sunstein et al. (1998) suggest that this is due not to differences in individuals' moral judgments or interpretations of fact situations, but rather to their difficulty in translating such judgments into financial terms.

3.5.2. *Effect of punitive damages on economic efficiency*

Only one study seeks to assess the effect of punitive damages on economic efficiency. Viscusi (1998) compares several measures of the costs of accidents from states that allow punitive damage awards to those from the four states that do not allow punitive damage awards (Michigan, Nebraska, New Hampshire, and Washington), controlling for other characteristics of states that may affect accident costs. He finds that allowing punitive damages has no significant effect on the rate of toxic chemical accidents, on toxic chemical releases, or on accidental death rates (but see the critique in Eisenberg, 1998). Because punitive damages have no effect on precautionary behavior, but impose substantial uncertainty and transaction costs on potential injurers, he concludes that they decrease efficiency.

3.6. *Mass torts*

How the liability system handles mass torts is another important policy issue that stretches across case types (see Hensler, 2001 for an excellent review and empirical analysis of mass torts, which we largely follow below). "Mass tort" is not a formal legal term; it usually denotes the consolidated litigation of a large number of tort claims arising out of a single accident or use of a single product. Such consolidation may include certification as a class action, but may also include other, less formal and less restrictive procedures (for discussion, see Peterson and Selvin, 1988). For example, "multi-districting," one doctrine used in mass tort cases, allows federal courts to aggregate cases involving common questions of fact into a single court by assigning responsibility for managing some or all phases of the litigation to a specific judge (Hensler et al., 1985). Empirically, mass torts account for only one-fifth of class action lawsuits, and class actions account for approximately 35 percent of mass torts (Hensler, 2001). In the 1980s, mass tort litigation grew significantly; Hensler and Peterson (1993) describe in detail the cases, including litigation over Agent Orange, Bendectin, the Dalkon Shield, the Hyatt skywalk collapse, the MGM Grand hotel fire, asbestos, and others, and provides some hypotheses about its causes.

As with class actions, the main efficiency argument in favor of the various mass tort procedural techniques is that they create economies of scale in litigation, thereby reducing the costs of evaluating numerous similar claims. No empirical studies have assessed comprehensively the efficiency or distributional consequences of mass-tort treatment of claims. However, in a number of formal reports and academic papers, researchers at the RAND Institute for Civil Justice have examined the characteristics of the largest and most costly (and still ongoing) of all mass torts—the asbestos litigation. Carroll et al. (2002) contains a comprehensive description of this work and bibliography; we summarize its key findings here.

Asbestos is a naturally-occurring mineral that is inexpensive to mine and process. Asbestos has many favorable properties, including strength, durability, and fire-retardancy. However, inhalation of asbestos fibers that occurs as a result of its handling and use causes a variety of lung diseases. Widespread occupational exposure to asbestos from 1940 to 1979 has been estimated to cause more than 225,000 premature deaths from 1985 though 2009, and large numbers of other less severe injuries and disability. After the 1973 decision in Fibreboard vs. Borel[8] finding asbestos manufacturers strictly liable to workers injured as a result of their product, increasing numbers of product liability lawsuits began to flow into the courts.

Carroll et al. (2002) report that over 600,000 people have filed asbestos claims through the end of 2000, with annual filings rising sharply in recent years. Over 6,000 companies in 75 out of 83 different types of industries have been named as defendants, ranging to include essentially every type of economic activity in the U.S. They find that a total of $54 billion has already been spent on asbestos litigation, over half of which has been consumed by transaction costs. Increasing numbers of less severe injuries account for most of the recent growth in the asbestos caseload. Estimates of the number of future claims and future costs vary widely. Best case scenarios predict approximately 1.2 million claimants and a total cost of $200 billion; worst case scenarios predict 3 million claimants and a total cost of $265 billion. The RAND researchers conclude that for asbestos compensation, the tort system is falling short of all of its principal objectives: deterrence of potential injurers, compensation of injured victims, and provision of "corrective justice."

4. Property

4.1. Introduction

The empirical literature relating to private property is quite diverse in its coverage. If there is one central theme, however, it relates to the design of mechanisms for allocating property rights and the rules for transfer of those rights, in the presence of transactions costs. The principal tradeoff behind these mechanisms is between incentives for optimal investment and transaction costs. This theme is not entirely new. Property law resembles contract law in situations where transaction costs are typically low, and resembles tort law in which transaction costs are often high.

In 4.2 we review studies of the development of private property; we follow this with an analysis of the efficiency/transaction costs tradeoff of different rules for the individual acquisition of previously shared property. Section 4.3 examines a closely related issue: the efficiency consequences of different rules of shared ownership—the allocation of rights to property for which the transaction costs of creation of individual rights is prohibitively high. In 4.4 we review the work that defines what the boundary of optimal property law should be, i.e., when property rules should be forgone in favor of liability rules. Work in this area was fundamentally shaped by Ronald Coase (1960), who

[8] 493 F.2d 1076, cert. denied 419 U.S. 869, 95 S.Ct 127, 42 L.Ed.2d 107.

observed that any allocation of property rights is efficient as long as there are no transaction costs and property rights are well-defined and enforced. Finally, 4.5 examines the consequences in terms of economic growth of the violation of the second assumption behind the Coase theorem: when property rights are poorly enforced and/or poorly defined.

4.2. The development of private property

The development of rights of private property has been a fundamental underpinning of our capitalist economic system. In the subsection that follows we discuss the creation of property rights generally; we follow this with a more focused discussion of property rights on shared property.

4.2.1. Creating property rights

Theoretical debates have surrounded such issues as the time span of private property rights, the alienability of those rights, and the appropriate use of public lands. Accompanying and complementing these theoretical debates have been a wide range of historical and empirical studies of the development and evolution of private property rights. For example, citing evidence from a study of Indians in Labrador, Demsetz (1967) posited the theory that all property rights regimes evolve efficiently in response to changes in economic conditions. Ellickson (1993) focuses his attention on regimes associated with close-knit societies in which power is broadly dispersed and members interact regularly. He questions the ability of Demsetz' theory to explain the development of the institution of slavery and other coercive regimes associated with cultures that were not close knit. With respect to close knit societies only, Ellickson points to economic efficiency, to the importance of non-economic concerns such as liberty, privacy, equality, and community. In support of his broader theory, he offers case studies of land ownership regimes at settlements during the colonial period in U.S history (Jamestown and Plymouth) as well as other non-U.S settlements, including those in Israeli (kibbutzim) and Mexico (ejidos).

A number of scholars have gone further back in time in their analysis of the evolution of private property, focusing on the development of rights of aboriginal peoples. Bailey (1992) reviews a wide range of studies of aboriginal societies, concluding that economic considerations play a significant role in explaining the evolution of rights. Moving from culture to culture, property rights vary according to use, resources, and particular economic circumstances. Bailey and others have emphasized the benefits associated with the development of private property rights.[9] Anderson and Swimmer (1997) broaden the discussion by emphasizing the importance of costs. They offer a more focused

[9] See, for example, Anderson and Hill (1990), who study the American West; Alston et al. (1996), who study the Brazilian frontier; and Anderson and Lueck (1992), who study American Indian reservations.

original study that includes a cross-sectional empirical analysis of 40 early American Indian tribes. Their analysis supports the conclusion that rights of access to property vary depend on the relative value of the property resources and the different costs of establishing and enforcing property rights.[10]

More modern studies of the evolution of property rights have focused on a number of closely related questions. First, can one explain which properties are owned publicly and which remain in the private sector; see, for example, Troesken (1997). Second, how important is the registration of land and the associated security of ownership important in the development of private property; see, for example, Miceli et al. (2001, Kenya), and Miceli and Sirmans (2000, U.S.). Third, what affect did local government regulations of land use have on the rate of development of private property; see, for example, Fischel (1990, growth controls).

4.2.2. Creating individual rights in shared property

One important deviation from the classic "Coase Theorem" framework arises when resources are "common pool." In such situations, access to a resource is at best only partly excludable; the resulting absence of entry controls leads to the "tragedy of the commons" in which the resource is overutilized. In essence, inefficiencies arise because there is no clear-cut mechanism by which individual agents can contract to exclude others and to prevent the dissipation or elimination of economic rents. Partial solutions to the tragedy of the commons involve means by which private property rights to common property are assigned. The empirical literature involves a series of qualitative studies, quantitative studies, and experiments that offer evidence as to the success of these private property regimes. Grafton et al. (2000) offer a relatively recent overview of the empirical literature, as well as a case study of a British Columbia fishery. Libecap (1989) describes the nature and complexities of the contracting issues that arise when one attempts to put into place property rights arrangements involving common resources.[11]

One empirical approach has been to compare the success of various private property regimes on a cross-sectional basis within the U.S. Agnello and Donnelley (1975) find that oyster fisheries in states with private property regimes were more productive than those in states that had limited open-access regimes. Townsend (1990) and De Alessi (1998) both point to evidence that the regulation of common resource fisheries is not enough; it is essential to create appropriate property rights. Grafton et al. (2000), focusing on halibut fishing, conclude that it may take years for the efficiency gains from privatization to materialize, and that both short-run and long-run gains can be at risk if there are restrictions on private property rights (on the duration, transferability, and divisibility of those rights and by other inhibiting regulations). They also point to the

[10] For further development of the theory underlie many of empirical studies, see, for example, Lueck (1994, 1995).

[11] For an application to the study of oil fields, see Libecap and Wiggins (1984); for parking, see Epstein (2001).

potential adverse impact of preexisting regulations, such as rate-of-return regulations on public utilities.

The particular method used to create private rights to public resources can obviously make a difference. The use of transferable property rights has become quite popular in recent years, both with restrict to the use of air resources (transferable emissions permits) and with respect to fisheries (transferable quotas).[12] Transferable permits have generally been deemed successful. With respect to sulfur dioxide, each public utility was given an allowance for tons of emissions each year; the private right is transferable and bankable. As reported by Schmalensee et al. (1998), the result was an active market in emissions rights from the inception of the program in March 1997, and an improvement in economic efficiency.

4.3. Shared ownership

Institutions can facilitate exchange by defining and enforcing property rights and by enforcing contractual agreements. One institutional form which can in theory achieve commercial benefits is the cooperative, an arrangement in which individuals share ownership of common property. Such ownership can be enforced purely through contract, or through the exercise of coercive power by the state. A number of historical studies have evaluated particular institutional arrangements to better understand the specific factors that affect the costs of creating and enforcing property rights. Pirrong (1995) reviews a variety of commodity exchanges, concluding that as a general rule commodity exchanges were able to facilitate private arrangements to share ownership. Specifically, the commodity exchanges were able to achieve five goals: commodity measurement, contract enforcement, the policing of theft and fraud, and the mitigation of information asymmetries. The one exception was the Chicago Board of Trade, which failed to regularize its measurement of grains and to reduce significant information asymmetries between buyers and sellers. This empirical study, along with a number of others, has served to provide support for the view of Ellickson (1991) and North (1990) that private cooperation can facilitate private rights without the state playing an active role. Ellickson's study of cattle farming in Shasta County, for example, leads one to conclude that norms of cooperation that are based on mutual interests of affected parties can serve as the basis of a system of shared ownership, even if the parties do not interact repeatedly.

4.4. Property rights versus liability rules

The extent to which rights of property ownership are valuable to individuals depends in part on the mechanisms by which these rights are enforced. In a classic article, Calabresi and Melamed (1972) characterize the trade-offs between liability rules (enforcement

[12] With respect to air resources (sulfur dioxide), see, for example, Joskow and Schmalensee (1998) and Joskow et al. (1998); with respect to fisheries, see, for example, Grafton et al. (1996).

through suits for damage) and property rules (enforcement through injunctive remedies). Property rule regimes are generally seen as more efficient when transactions costs and the likelihood of strategic behavior are low, whereas liability rule regimes can be designed to minimize the possibility of strategic behavior and can work more effectively when transactions costs are high.[13]

A variety of empirical and experimental studies have been undertaken to evaluate the extent to which the law relies on one or both enforcement mechanisms and the effectiveness of those mechanisms. One set of applications has included empirical studies of the extent to which bargaining arises in property rights situations. For example, Farnsworth (1999) finds no evidence that parties bargain after judgments are reached in nuisance cases; this counters one's expectations that a judgment would serve to clarify the assignment of property rights. Apparently in the 20 nuisance cases under study the relationship among the parties was sufficiently acrimonious as to make bargaining difficult.

Another set of applications of considerable interest has been the normative and positive evaluation of situations under which the government has required the sale of private property. Thus, a regulatory taking can be seen as the imposition of a liability rule to enforce the ownership of certain types of private property, whereas property otherwise immune from a taking would be protected by a property rule.

While the takings debate has generated substantial normative debate, the empirical analyses of the pros and cons of government regulation of private property have been much more limited. A 1998 paper by Heller (1998) focuses on the "anti-commons problem." In an anti-commons, multiple owners of a property each have the right to exclude others from the use of that resource, but no one has the privilege of sole use of the property. With too many owners exercising their right to exclude, the property is likely to be under- rather than over-used. Heller applies this theory to the utilization of property rights in Russia, explaining, for example, why many Russian storefronts were empty while street kiosks were full of goods. In an interesting empirical study, Miceli and Sirmans (2000) point out that even if it appears efficient to divide a property; that such a situation is likely to generate hold-out problems. They evaluate conditions under which the imposition of a "right to partition" the land, can (and cannot) improve social welfare. The authors also evaluate a remedy that requires the forced sale of the parcel of land, with the proceeds divided among the owners. They find that this remedy is often unfair and inefficient because it does not account for and preserve the subjective values that individuals may place on their particular parcels of land. They prefer a remedy in which the land is partitioned, and individual property ownership is maintained, despite the fact that there is a risk of strategic behavior and a potential cost of buying and reassembling the land.

[13] Calabresi and Melamed (1972) consider a third method for enforcing property rights—inalienability (rights that cannot be given away or sold, such including the vote, human rights, and certain body organs. For a valuable discussion of the conditions under which alienability applies, see Rose-Ackerman (1985).

In one intriguing study of the evolution of agricultural and fishing rights in Iceland, Eggertsson (1992) describes how initial private ownership of large tracts of land evolved into communal property. The resulting commons problem was responsible, in part (other causes included adverse trade relations, volcanic eruptions, pests and plagues, and a cooling climate) for the long-term economic decline of the Icelandic economy, running through the 18th century.

4.5. Consequences for efficiency and economic growth of poorly defined/enforced property rights

A system of poorly defined or inadequately enforced property rights can be expected, as a matter of theory to create economic inefficiencies and to inhibit or otherwise limit economic growth. In short, economic growth requires a government or governmental entities that establish property rights so individuals and firms can contract with a minimum of transactions costs.[14] A variety of historical studies, covering a range of substantive areas, provide support for this proposition.[15] With respect to water rights, the changes that occurred in the Western U.S. in the latter part of the 19th century offer a case in point. Kanazawa (1998) points out that California had adopted a system of water law that recognizes both riparian and appropriative rights. In riparian regimes rights are based on ownership of land and water rights are derived from use, not from physical possession. These rights are often less clear than appropriate rights, which are based on possession. Kanazawa offers empirical evidence that courts tended to promote appropriative claims when transactions costs were high to encourage the reallocation of water rights to their highest and best use.

Another focus in the property rights literature has been on the question of the extent to which property rights are essential if one is to improve the well-being of the poorest nations and their inhabitants. Using a cross-section sample of 47 countries with data focusing on changes from 1960 to 1989, Keefer and Knack (1997) find that deficient property rights institutions (as measured by indicators of the rules of law, the pervasiveness of corruption, and the risk of expropriation and contract repudiation) cause poor countries to fall back rather than catch up with their richer counterparts. More recently, using a cross-section of over 100 countries for the 1975–95 period, Norton (1998) finds that well-defined property rights improve the lot of both poorer countries and the poorest inhabitants of those countries.[16]

In recent years, it has become more and more apparently that the specification and effective enforcement of a system of private property rights is essential if developing economies are to make a successful transition to a democratic capitalistic system which

[14] The underlying theory is laid out, for example, in Norton (1998) and North (1981).

[15] These studies include work by Barro (1991), Keefer and Knack (1997), and Knack and Keefer (1995).

[16] Other recent studies with a geographical focus include Migot-Adholla et al. (1991; Sub-Sahara Africa) and Lanjouw and Levy (2002; Ecuador). For a study of the importance of the effectiveness of government, see, for example, La Porta et al. (1999).

supports economic growth. For an interesting discussion of the difficulties of achieving a transition to a system of private property in Russia, see, for example, Shleifer, Boycko, and Vishny (1993, 1995).

5. The litigation process

5.1. Introduction: the costs of litigation

Litigation arises when the action of an injurer harms or allegedly harms a victim. The harm may be intentional as in the case of most crimes, accidental as in the case of many torts, or incidental as in the case of many nuisance situations. The behavior of injurers, and, in some cases, victims (their precaution and levels of activity) affect the frequency and extent of harm to one or both parties. Economic efficiency requires balancing the cost of harm against the cost of avoiding it, including the costs of risk bearing.

Most early studies of the costs of litigation were longitudinal in nature. In conducting those studies, social scientists focused primarily on the "supply" of social harm.[17] In this framework, harms are seen not only as a function of the particularities of the legal system, but also as the consequence of more basic socioeconomic determinants. For example, automobile accident litigation may be affected by automobile registrations, gasoline prices, the average age of drivers, income levels, and urbanization. With respect to the legal system, differences among legal rules (e.g., strict liability vs. negligence), differences in rights of action, whether the trier-of-fact is a jury or a judge, and whether the attorney works on an hourly or contingent fee basis, are among the factors that can affect the costs of litigation.

Clark's (1990) comparison of litigation trends in Europe and Latin America since 1945 is a useful example of a longitudinal study of litigation. Clark suggests different rates of litigation among Spanish regions from 1945 to 1967 or for Italian regions from 1952 to 1968 are best viewed as representing differential adjustments to varying environmental conditions. However, Clark and other similar longitudinal studies tend to avoid structural explanations that might reflect underlying sociological, economic, and political factors. Other than a passing reference to theories of the business cycle and to models of the evolution of courts, Clark does not look for variables that might characterize changes in the economy or to changes in the legal system that would produce more or less harmful behavior.[18]

More recent time series studies of litigation have focused on the apparent "explosion" of litigation that occurred in the 1960s and 1970s. Analyses of these and more recent periods have typically been descriptive in style, emphasizing the substantive areas of law

[17] See Cooter and Rubinfeld (1990).

[18] A conceptually more appealing analysis is that of Stookey (1990), whose model allows social crises to affect rates of dispute, which in turn affect litigation rats. Stookey notes that changes in litigation rates crate pressures for policy changes, which in turn induce changes in litigation rates.

in which litigation has grown most rapidly. For example, Galanter (1986) describes a rapid increase in products liability filings and in jury awards. Galanter also points to the fact that more than 98% of all civil cases are filing in state rather than federal court. As a result, any significant trends in litigation rates are likely to show up using state court data, which has been reported over the years by the National Center for State Courts. Finally, Galanter points to the interesting comparative fact that while many countries have lower overall rates of case filings than the U.S., that the per-capita use of the courts is in the same range in the U.S., Canada, Australia, England, Denmark, and Israel.[19]

5.2. The chronology of a legal suit

Initially, in the *first stage* of the chronology of a lawsuit, there is an underlying event in which one or more persons or entities (an "injurer" or "injurers") allegedly harms others (a "victim" or "victims"). The frequency of harm will generally be influenced by the choices that potential injurers make about the levels of activity in which they engage, and the precaution that they take when engaging in those activities. Activity levels are typically not easily monitored by public enforcers or by courts, so that much litigation (and the empirical evidence surrounding that litigation) centers on the alleged injurer's level of precaution. The activity levels and extent of precaution chosen by those who are potential victims may also affect the frequency of harm.

As will be true at each stage of the litigation "game," the "rational" decisions of potential injurers and victims will be dependent on their expectations as to the stream of benefits and costs associated with their activity and precaution choices. These benefit and cost streams, are of course, related directly to the decisions that are likely to be made in each of the further stages of the litigation game.[20]

In the *second stage* of a legal dispute, the party that allegedly suffered harm typically hires an attorney and then decides whether and where to assert a claim. If made on a rational and self-interested basis, these decisions will be the solution to a sequential game that balances the costs of asserting the claim against the benefits that might be expected through settlement or trial. The extent to which the attorney's interests are or are not aligned with those of the client is a significant issue for any legal system. Moreover, this principal-agent problem is likely to be significantly influenced by the fee arrangement under which the attorney operates.

The *third stage* of litigation occurs after the claim has been filed, but before trial. Pre-trial litigation efforts, which relate generally to the discovery process, include responding to complaints, answering interrogatories, and taking depositions. By providing information, the discovery process serves to help the parties sharpen their expectations about trial outcomes. Consequently, the pre-trial discovery process provides the central

[19] Galanter (1986), p. 7.

[20] For a general overview of the litigation process, with an emphasis on normative issues, see Cooter and Rubinfeld (1989); see also Kaplow and Shavell (2002).

input for the *fourth stage* of the litigation process, settlement bargaining. If the parties can reach a cooperative outcome is reached the case settles; the non-cooperative alternative is a trial.

The *fifth* stage of the litigation process is the trial itself. Trials can be costly exercises, although costs are likely to vary widely depending on the stakes in the case, and the nature of the forum in which the trial takes place (jury vs. judge trial, arbitration vs. litigation, etc.). Because litigation is costly, trials are typically negative sum games, with a possible exception that can arise when external benefits to one party make the payoffs of the trial game substantially asymmetric.

It is useful for our purposes to also distinguish a *sixth stage* of the litigation game—the trial outcome itself. Judgments by judges or juries can be injunctive or compensatory or a combination of the two. If both parties are satisfied with the judgment, the litigation process ends. However, one or both parties can opt for the *seventh stage*—an appeal to a higher court or courts.

The theoretical game-theoretic literature describing each of the stages of the litigation process, and/or combinations of several stages is substantial. However, the empirical evidence that is needed to provide the foundation for those models is in much more limited supply. In the subsections that follow, we highlight some of the relevant empirical literature. Because our focus is on the litigation process itself, we begin with the filing of the lawsuit.

5.3. Forum and law choice

An individual who believes that he or she has suffered harm must (with the assistance of an attorney) choose the appropriate law and the appropriate forum in which to file. It is not unusual for the victim to have a range of law and forum options, depending on whether state, federal, or foreign statutes apply. Moreover, defendants face forum choice options as well. Thus, a defendant facing a patient infringement claim in a jurisdiction believed to be sympathetic to the plaintiff may choose to file a declaratory judgment claim in a more friendly location.

The ability and the incentive to forum shop are both likely to be influenced by the decision rules that courts use to resolve conflicting forum claims. There are two related questions of interest. First, suppose that a claim is filed in one jurisdiction involving parties that operate in several jurisdictions. Which jurisdiction's law should apply? Second, suppose that the parties file related claims in two distinct jurisdictions. How do the jurisdictions decide which should take precedence?

The former question has been the focus of much legal debate and some empirical analysis. Indeed, the legal footing on which the choice of law (or alternatively the conflict of laws) doctrine relies has evolved over the past half century. Under the First Restatement of Conflict of Laws (1934), for example, solutions were likely to depend on non-legal factors, rather than on the substantive laws of the conflicting jurisdictions. The Second Restatement (1971) gives credence to the laws of the possible forums, but suggests taking into account a wide range of factors, including, but not limited to: (i) the relevant policies of each forum; (ii) the protection of justified expectations; (iii) the certainty, predictability, and uniformity of results; and (iv) the ease of determination and application of the law.[21]

Under the Second Restatement, judges are afforded substantial discretion in making choice of law decisions. In a study of 802 multi-court opinions involving choice of laws, Borchers (1992) found a wide variation in the pattern of results. Courts that tended to rely on the First Restatement tended to favor rules that favored plaintiffs' recovery less often than those that relied on the Second Restatement. In addition, courts that relied on the First Restatement were less likely to benefit local parties than those than did not. On the whole, Borchers reaches the conclusion that whatever the cited basis, most courts use their own judgment, in the process finding a way to do what they believe is substantively appropriate without necessarily following the law. However, Thiel (2000) comes to a contrary conclusion, based on a reanalysis of Borchers' data using more sophisticated econometric techniques.[22] Thiel finds little evidence to support the theory that judges are more favorably inclined to residents or to plaintiffs, although he does (not surprisingly) find that judges tend to favor the law of their own forum.

Thiel's article is (to our knowledge) the first that makes an effort to draw empirical implications from theories of interest-group politics. He suggests that the competition among jurisdictional laws can be seen as dynamic prisoners' dilemma game, in which the cooperative outcome (which maximizes the joint welfare of the two jurisdictions) may or may not be achieved.[23]

Whether forum shopping enhances or diminishes economic efficiency is open to debate. The positive case for shopping is based on the possibility that a party can choose a forum that will allow from a quicker and more accurate trial. For example, a number of authors have argued that Delaware's successful effort to develop a specialization as a center for corporate law has generated a positive "race-to-the-top," not a race-to-the bottom.[24] The case against forum shopping is based on the view that differences in expected awards across forums reflect true differences in the cost of failing to take precaution. Suppose, for example, that the cost of precaution is sufficiently high in the jurisdiction in which a plaintiff is injured so as to make the injurer non-negligent in tort. If the plaintiff were to file in a jurisdiction in which precaution is less costly, the defendant will be overdeterred and litigation costs will be excessive.[25]

There is evidence that forum shopping has increased substantially over the past thirty years. For example, the fraction of state court cases that were removed to federal court

[21] See Borchers (1992), p. 362, citing section 6 of the Second Restatement.

[22] Thiel uses a two-limit Tobit model, which builds on the assumptions that (1) some choice of law arguments are not made by each party to the litigation because the interest is not sufficient to cross an appropriate threshold, and (2) judges can only apply choice of law analyzes when the issue is raised by one or both parties.

[23] For another study that takes an interest-group, political science perspective, see Solimine (1989).

[24] See, for example, Bebchuk et al. (2002) and Romano (2002).

[25] For a normative discussion of forum shopping along with empirical evidence, see Clermont and Eisenberg (1995).

increased from 15% in 1970 to over 30% by the year 2000.[26] This is consistent with the view that defendants often choose more friendly forums after plaintiffs have filed suit. The alternative explanation is that defendants are simply responding to an inefficient forum choice by plaintiff. Clermont and Eisenberg recently attempted to sort out a number of competing interpretations of the increased forum shopping.[27] They studied a large sample of cases (utilizing a database obtained from the Administrative Office of the Courts of three million federal cases terminated over thirteen years) involving diversity of jurisdiction; in such cases federal as well as state court venues are available to the parties. While plaintiffs were successful in 71% of cases overall, their success in cases removed to federal court was only 34%. This was due in part to the ability of defendants to move cases away from plaintiff-friendly jurisdictions (which may or may not be efficiency-enhancing). However, it also reflects a case selection effect, since to some extent the cases that were removed were relatively weak from plaintiff's perspective, wherein the strong plaintiffs' cases had been settled or otherwise resolved. Using regression analysis to control for characteristics of cases and hopefully the case-selection effect, the authors concluded that forum-shopping effects remain substantial. Forum selection through removal, all else the same, reduces a plaintiff's probability of success from 50% to 39%.[28]

Because of their focus on diversity cases, Clermont and Eisenberg make the plausible presumption that the dominant advantage of forum shopping inures to the defendant. Their view is that the transfer of forum causes some plaintiffs to abandon their cases or to settle on relatively unfavorable terms; moreover, it also makes litigation more difficult for the plaintiff. Yet, absent the ability to transfer, forum shopping clearly can and has benefited plaintiffs. One recent study lays out the substantial benefits to both parties from forum shopping in patent litigation cases.[29]

In summary, empirical evidence supports the view that there is substantial forum shopping by plaintiffs, whose effects are diminished but not eliminated by defendants' ability to transfer cases. These effects appear significant, even after controlling for important case and selection effects. As with other areas of litigation, there remain many unanswered questions related to the choice of laws and forum. To what extent do defendants make an initial strategic forum choice, as, for example, by filing for declaratory judgment? What are transactions costs that are imposed by the forum-shopping process? What risks are put on the parties by the uncertainty surrounding choice of law and choice of forum? While the current state of empirical research is useful, more will be needed before the normative issues relating to forum and law choice can be resolved.

[26] Clermont and Eisenberg (2002), p. 4.

[27] This discussion is based on Clermont and Eisenberg (1998), which reports results from their 1995 paper.

[28] The authors also examined cases that were transferred from court to another, rather than removed from state to federal court. Plaintiffs' probability of winning in transferred cases, other things equal, falls from 50% to 40%.

[29] Moore (2001).

5.4. Pre-trial

The litigation process is best viewed as a non-zero sum game in which each of the parties reveals information about the strength of its strategic threat point. It is widely accepted that trials tend to arise when the judgment expected by the plaintiff is greater than the judgment expected by the defendant, i.e., when the plaintiff is *relatively optimistic*. Indeed, in order for settlement to be possible, the savings in costs to the parties must exceed their relative optimism about trial.[30] Thus, settlements tend to occur when relative optimism is small, whereas trials tend to occur when the plaintiff expects a large trial judgment and the defendant expects a relatively small judgment. Typically, one would expect the pre-trial discovery process to move the discovering party's expectation closer to the trial outcome that would occur if all information were pooled, while at the same time reducing the variance of the distribution of expected values, thereby reducing the risks that the litigating parties face.

Any pre-trial information that decreases the mean of the plaintiff's distribution of possible trial outcomes will make the plaintiff less optimistic and thereby decrease the probability of trial, and conversely for the defendant. Moreover, because trial occurs in a relatively small number of cases where the plaintiff is optimistic relative to the defendant, a reduction in variance caused by the pooling of information is likely to result in fewer trials as well. It is useful to view the pre-trial discovery process as including the voluntary disclosure of information, which will tend to support each side's optimism, and the mandatory disclosure, which will have the opposite tendency.[31]

Pre-trial discovery can satisfy an additional goal, by increasing the likely accuracy of any settlement that the parties reach. An accurate settlement is a hypothetical settlement that would be reached if the parties had complete information about the law and the facts of the case, and the defendant paid an amount equal to the complete information judgment to the plaintiff (or a correspondingly appropriate injunction were imposed if an equitable remedy were required). Similarly, discovery can increase the accuracy of trials by improving the information available to the court.

Finally, the legal rules surrounding pre-trial processes also have implications for the incentive of the parties to take precaution and on the probability that harm will occur in the first place. Discovery that is likely to increase the relative optimism of the plaintiff (or the relative pessimism of the defendant) is likely to increased the expected value of a claim, and thereby the likelihood that a case will be filed.[32] This will, in turn, increase deterrence by increasing the potential defendant's incentive to take precaution. Discovery could encourage plaintiffs to drop claims that have little merit, i.e., claims that were

[30] For an overview of the economics of discovery, see Rubinfeld (1998) on which portions of this subsection are based; for a more formal analysis, see Cooter and Rubinfeld (1994).

[31] For a general discussion of voluntary vs. mandatory disclosure see Hay (1994) and Shavell (1989).

[32] This abstracts from changes in the probability of settlement; if plaintiff becomes more optimistic, the probability of trial could increase, which, other things the same, could diminish the expected value of the claim (because trial costs could be large).

based on false optimism. In general, broad discovery rules are likely to increase deterrence to the extent that they improve the specificity of the litigation process, which has the effect of increasing the gap in liability costs between failing to take care and taking care.[33]

The theory of pre-trial discovery makes clear the trade off between the costs of pretrial discovery and the costs of trial. A more extensive and more costly discovery process is likely to increase the probability of settlement (which is typically well over 90%) and thus to save on trial costs. However, discovery can itself be highly costly, and the appropriate balance is not clear. If discovery had no effect on settlement, the cost of obtaining information through discovery should be around 10% of the cost of finding that information through trial in order for discovery to save transactions costs directly. However, discovery can increase the likely of settlement, which itself reduces transactions costs, so the ultimate trade-off is less transparent. Moreover, discovery can be used strategically by the parties; since discovery costs are typically borne by the responding party, it can be advantage for a party requesting discovery to engage in broad discovery simply to raise its rival's costs.

Interestingly, there is substantial variation in discovery practices across different legal environments, which offers a potential and relatively untapped source of empirical information that is relevant to the normative questions that have just been asked. We know that discovery practices (and costs) vary with legal environments, based among other things, on (i) whether the system is adversarial as in the U.S. (a common law country), or inquisitional, as in most European countries (which operate under civil law); (ii) whether there is substantial reliance on juries as triers-of-fact, as in the U.S., but not England[34]; (iii) whether there is substantial fee shifting from losing to winning parties; as in the U.K. and parts of Europe, but not in the U.S.; and (iv) whether the underlying culture relies heavily on attorneys, and relies on the court system to encourage the provision of information and to resolve disputes, as in the U.S., but not in Japan.[35]

Some historical, primarily non-econometric evidence supports the view that discovery efforts are responsive to economic incentives. Thus, there is a strong finding of a positive correlation between the stakes of the case and discovery effort.[36] Surprisingly, however, one study found that cases with greater discovery tended to settle less than those with less discovery.[37] While the author did control for case complexity and case size, the formal analysis did not reflect the relatively complex set of strategic incentives facing the parties.

[33] P'ng (1987).

[34] There is some empirical evidence that delays in getting to trial are on average greater for judge trials than for jury trials (Eisenberg and Clermont, 1996), and alternative dispute resolution processes do not reduce delay (Heise, 2000).

[35] Langbein (1985) offers an interesting qualitative comparison of discovery in the U.S. and in Germany. For a comparison with Japan, see Ramseyer and Rasmusen (1997).

[36] See, for example, Trubek et al. (1983), pp. 95–96 and Glaser (1968). For a more general overview of the non-econometric empirical evidence, see McKenna and Wiggins (1998).

[37] Glazer (1968).

On the whole, there has been very little econometric work on the pre-trial behavior of the parties. One important exception is Shepherd (1999). In a study based on an interview survey of attorneys in 369 federal civil cases, Shepherd estimated a simultaneous-equations tobit model that emphasized the interdependent nature of the strategic discovery choices made by plaintiffs and defendants. He modeled each party's discovery effort as a function of the expected recovery in the case, the expected non-monetary recovery, various case-specific factors, the fee-arrangement with the client, and the discovery effort by the opposing party. Shepherd found that plaintiffs typically choose their discovery independently of the defendant, i.e., they do not respond strategically. Defendants, however, respond almost entirely to the plaintiffs' discovery efforts in a manner consistent with a tit-for-tag game-theoretic strategy. Finally, plaintiff attorneys operating under hourly fee arrangements tend to conduct on average more than 5 days more discovery than do attorneys operating under a contingency. Whether these and other results will hold up in future studies that use data sets that allow one control more effectively for case characteristics and to measure discovery effort objectively remains an open question.

To our knowledge, there has been little or no research into the costs imposed on the larger judicial system by the discovery process. Thus, we know little about the costs to the parties that could associated with the revelation of confidential information, or the costs to non-litigants who might be affected by case outcomes. Nor, do we have much information on "discovery abuse," which might be defined as discovery requests where the direct information benefit to the requesting party is less than the cost to both parties of meeting that request.

Much of the empirical analysis of pre-trial discovery has focused on the benefits and costs of changes in discovery rules, with the most recent evidence flowing from a debate as to whether recent changes in federal discovery rules that required substantial disclosure by both parties.[38] Because half of the local districts opting out of the required changes, as allowed under the reforms, a more or less natural experiment (but-for selection effects) was created. Key questions were (i) would mandatory disclosure increase or decreasing the costs of litigation, and, (ii) would mandatory disclosure delay or speed up the litigation process?

One recent study by the Federal Judicial Center surveyed 2,000 attorneys involved in 1,000 cases that were terminated in 1996.[39] The survey suggested that attorneys, on balance, believed that the accuracy of the process was improved, which would tend to lower costs. Moreover, there was evidence of a small decrease in the time involved in litigating cases. A second study, conducted by the RAND Corporation, compared the behavior of a small group of district courts prior to the discovery reforms, some of which

[38] The majority of studies of the role of discovery have been undertaken primarily by socio-legal scholars. See, for example, a Federal Judicial Center study (Connolly et al., 1978), the work that emanated from the Wisconsin Civil Litigation Research Project (Trubek et al., 1983), or more recently a study by the National Center for State Courts (Keilitz et al., 1993).

[39] Willging et al. (1997).

had a form of mandatory disclosure and some of which did not.[40] In this case, the RAND group found no significant differences in delay or litigation costs, but some increased dissatisfaction of attorneys who operated under mandatory disclosure rules. Whether changes in disclosure rules really have little effect, or whether we simply have not been able to measure those effects with sufficient precision, remains an open question.[41]

5.5. Settlement

We suggested previously that cases are less likely to settle the greater the relative optimism of the plaintiff about the trial outcome. More generally, the basic economic model of litigation suggests that settlements will occur if there is a positive cooperative surplus to be enjoyed by settling.[42] This surplus consists of three elements (1) the sum of the trial costs of the two parties; (2) the difference between the plaintiff's and defendant's subjective expected payoffs from going to trial; and (3) the sum of the risk-bearing costs borne by both parties. To put this formally, let

$T_{\mathrm{p}}(T_{\mathrm{d}}) =$ subjective value to plaintiff (defendant) associated with a successful trial outcome;

$c_{\mathrm{tp}}(c_{\mathrm{td}}) =$ cost to plaintiff if the case is settled.

Then, according to theory, cases will settle when $G = (T_{\mathrm{p}} - T_{\mathrm{d}}) + (c_{\mathrm{tp}} + c_{\mathrm{td}}) > 0$.

This model of settlement rules out the possibility of strategic behavior by one or both parties; such behavior must be the result of an effort by the parties to increase their share of the cooperative surplus from settling the case. One practical approach to account for strategic behavior is to add randomness to the settlement-trial decision, in which case the probability of settlement is a function (usually a cumulative distribution function) of the size of the cooperative surplus, measured directly or proxied by a series of case-specific variables. A substantial number of empirical studies of the settlement-trial decision have been authored over the years. The earliest studies by socio-legal scholars have been both longitudinal and cross-sectional. Only in the past decade or two, however, have economists begun the task of specifying and estimating structural models of the behavior of the parties, and they have done so to answer a variety of interesting questions.

One question is whether augmenting compensatory damages will result in more or fewer trials. Theory tells us to expected three competing effects: (1) the tendency of relative optimism to cause cases to go to trial will be strengthened, which will reduce the probability of settlement; (2) the cost of trying cases will increase, which will increase the probability of settlement; and (3) the risks faced by both parties will increase, which

[40] Kakalik et al. (1998).

[41] For another study that found no effects, see Huang (2000).

[42] For overall reviews of the relevant theoretical literature, see Cooter and Rubinfeld (1989), Daugherty (2000), Hay and Spier (1998), and Kaplow and Shavell (2002).

should increase the probability of settlement. In one study of a large number of antitrust trials, Perloff and Rubinfeld (1987) found evidence suggesting that, where reputation effects are important and where the parties tend to be pessimistic, treble damages leads to a decrease in the proportion of settled cases and an increase in the number of trials.

Other studies, however, have come to contrary conclusions, including Danzon and Lillard (1983), who studied medical malpractice claims. Danzon and Lillard model four equations: two trial equations explain the probability of plaintiff winning and the amount of the verdict, while the settlement equations explain the minimum demand of the plaintiff and the maximum offer of the defendant. The authors assume that increased stakes increase the randomness in the model proportionally, but the cost of litigation less than proportionately. This leads to their conclusion that higher stakes will cause more cases to be litigated. They also find the higher the plaintiff's probability of winning at trial the lower the probability that the case will be tried.

Recent studies of settlement behavior have become more sophisticated in their ability to account for selection effects and the close link between settlement behavior and trial outcomes. Viscusi (1988) focuses on the decisions to drop and to settle product liability claims. He concludes that the expected award had from two to nine times the influence of the variance on settlements. Fournier and Zuehlke (1989) estimate a plaintiff's settlement demand equation that explicitly accounts for self-selection, using a survey of civil filings from 1979–81. They find that higher trial awards and increased variance both increase the probability of settlement.

Viscusi and Scharff (1996) account for selection effects at both the drop and the settlement stages of litigation using a large number of product liability cases. They find that higher trial awards have a greater (positive) effect on the defendant's settlement offer than on the plaintiff's request. Conversely, higher probabilities of winning at trial have a greater effect on the plaintiff's request than the defendant's offer. The first result is consistent with a defendant that is more risk averse than a plaintiff, whereas the second result suggest the converse.

Using the same data set described previously, Perloff, Rubinfeld, and Ruud (1996) estimated a joint model of trial outcomes and settlements. They modeled settlements as a function of the expected distribution of outcomes, conditional on a case going to trial. In their model the probability of settlement is a function of the probability that plaintiff and defendant expect to win at trial as well as the variance of the trial outcome. They find that risk aversion plays an important role in explaining why cases settle instead of going to trial. When the probability of winning is near 50%, the variance of trial outcomes is at its maximum.[43] When the probability of success increases above 50%, the variance decreases. Using this fact, and accounting for the importance of risk aversion, the authors find that for every 1% increase in the probability that the plaintiff wins, the probability that the case settles increases by nearly 0.13%. Moreover, because

[43] In a binomial distribution with success probability p, the variance is $p(1-p)$, which is maximized at $p = 1/2$.

the magnitude of the risk aversion effect increase with the size of damages awarded, trebling antitrust damages has a dramatic effect on the probability of settlement. Once again, were treble damages eliminated, the fraction of cases going to trial would increase substantially.

While reputation effects per se may or may not be significant in particular cases, the likelihood that settlement will be reached may be a function of the nature of the interactions among parties. In one recent study Johnston and Waldfogel (2002) found even when litigants have no history, that attorneys may be involved in repeated play. Using a data set on 2,000 federal civil cases filed in the Eastern District of Pennsylvania in 1994, they found that a history of attorney repeat interaction has a significant positive effect on the probability that cases settle and that such interactions tend to shorten the disposition of the litigation generally.

The particulars of settlement behavior depend on the nature of the parties (firms or individuals, risk neutral or risk-averse, etc.), on the nature of the cases (large stakes, small stakes, reputation effects, etc.), and more generally on the institutional characteristics associated with the particular subject matter at issue. We have cited previously a range of studies of civil litigation generally, and of torts, antitrust, products liability and medical malpractice.[44] However, the range of cases covered almost all areas of civil litigation, including tax and securities.[45]

Settlement is often seen as the preferred outcome from a policy perspective because it avoids trial costs. However, seen from the perspective of the litigants themselves, the preference for settlement is not so clear. On one hand, Gross and Syverud (1996) explain, relying in part on a survey of attorneys, the pervasive preference for settlement in most legal systems.[46] On the other hand, Galanter and Cahill (1994) point out that efforts by the courts to encourage settlement, such as court-ordered mediation, are met with mixed satisfaction by the litigants.[47]

5.6. Trial

The outcome at trial is the result of a complex interaction between the efforts that both parties put into the trial and the underlying facts and law of the case. Perhaps the most central topic of empirical analysis has been the fraction of cases that are won by plaintiffs and by defendants at trial. All researchers working on this question are aware that

[44] Other relevant medical malpractice studies include Hughes (1989), Sloan and Hoerger (1991), and Farber and White (1991).

[45] Lederman (1999) studied 400 Tax Court cases, and found significant selection effects. Alexander (1991) found that fewer than five percent of litigated securities cases were tried to judgment.

[46] See also Gross and Syverud (1991).

[47] Institutional constraints can affect settlement rates as well. Based on an analysis of the timing of settlements of automobile bodily injury insurance claims, Kessler (1996) found that delay in the trial court system generally increases delay in settlements, and that tort reforms designed to reduce delay in settlement did not work as intended. Kessler used two separate cross-sections of settled claims, one from 1977 and one from 1987.

the concepts of winning and losing are not well defined, since both may be relative rather than absolute concepts. The plaintiff "wins" a civil suit, in one sense, if the court awards damages or provides injunctive relief. Many civil suits, however, concern not the fact of defendant's liability, but its extent. From this perspective, the plaintiff "wins" at trial only if the damage award is larger than the defendant's settlement offer.

In any case, it is well known that the selection of cases for trial need not, indeed, is unlikely to be, a random selection of filed cases. Despite the former, and in light of the latter, a number of papers have focused on whether and under what conditions one should expect plaintiffs to have a 50% win rate. Priest and Klein (1984) have received prominent treatment in this debate, in part because they were careful to think through the selection-bias issues that arise when one accounts for the fact that those cases that go to trial are likely to be those for which there is no (positive) cooperative surplus. Their starting point is that one should expect a 50% win rate; if the rate tended to be higher, for example, defendants would find it in their advantage to settle weaker cases, and vice versa.

Others, however, have pointed out that the interests of the parties and their attitudes towards risk may not be symmetric, which could generate win rates that are significantly different from 50% for plaintiff. For example, if one or both parties are concerned about reputation effects, the benefits of winning may be greater than those that are trial specific. Arguing that defendants are likely to be more concerned with reputation than plaintiffs, Perloff and Rubinfeld (1987) found that approximately 70% of all antitrust cases in their data set were won by defendants. Wittman (1985, 1988) argued further than when the parties disagree about the expected trial award, win rates can deviate substantially from 50%. Perhaps the general points here are best summarized by Shavell (1996), who points out it is possible for the cases that go to trial to result in plaintiff victory with any probability, and consequently, that the probability of plaintiff victory in settled cases, had they been tried, may be any other probability.

Waldfogel (1995) has offered a thorough and broad empirical analysis of the selection of cases for trial.[48] Using data describing contracts, intellectual property, and torts cases filed in the Southern District of New York from 1984–87 and closed by 1989, Waldfogel finds strong statistical relationships between the rate at which cases go to trial and plaintiff win rates. When relatively few cases go to trial, the plaintiff win rate tends to be close to 50% as predicted by the model of selection bias. However, when the trial rate increases, which is presumably due to significant asymmetries between plaintiff and defendant expectations about trial outcomes, the plaintiff's success rates diverging from 50%, with rates being higher in some cases and lower in others. Because of data limitations, Waldfogel is unable to explain the underlying source of variation in win rates across case types.[49]

[48] For an earlier analysis of similar cases, see Eisenberg (1990).

[49] See also Siegelman and Donohue (1995), who discuss the complexities of sorting out settlement decisions and trial outcomes in the context of employment disputes.

In a somewhat more recent paper, Waldfogel (1998) uses the same data set to evaluate the divergent expectations explanation of why cases go to trial (as discussed previously) with an alternative asymmetric information theory. Under the latter theory one party, perhaps the defendant, knows the probability that plaintiff will win at trial, while plaintiff knows only the distribution of plaintiff victory probabilities. If the uninformed plaintiff makes a settlement demand, the defendant will tend to accept that demand only if the plaintiff's probability of success is relatively high. As a result, for those cases proceeding to trial, the plaintiff's win rate should be consistently below the fraction of plaintiff winners in the pool of all cases, and presumably smaller than 50%. Waldfogel's evidence rejects the asymmetric information theory in favor of the divergent expectations theory discussed previously.

There is little doubt, as Eisenberg and Farber (1997) point out, that the process by which lawsuits get filed and go to trial depends not only on the subjective expected value of the claim, but also psychological and other non-pecuniary motives. In a broad sense, those cases that are filed and that eventually go to trial, are likely, other things equal, to be those where the costs of litigation are lowest. Interpreting costs broadly, we can say that a "litigious" plaintiff is one that views the costs of going to trial to be unusually low. Using a data set on over 200,000 federal civil litigations, the authors find that plaintiffs who are individuals are likely to have higher rates of trial than corporations (lower settlement rates), and consistent with the implied asymmetry of interests, lower plaintiff win rates.[50]

The costs of litigation are in part a positive function of the time it takes to dispose of a case at trial or through settlement. A number of economists and socio-legal scholars have studied the factors that affect the time of disposition of civil trials. One important effort along these lines has been made by researchers at the RAND Corporation's Institute for Civil Justice. In evaluating the Civil Justice Reform Act of 1990, RAND studied one year of civil jury case outcomes from 45 of the 75 most populous counties in the U.S. The study found that the average disposition time for a case that went to trial in Cook County Illinois, was over five years.[51]

A useful overview of trends in civil jury verdicts for the decade beginning in 1985 is offered by the Moller (1996) RAND study. Focusing on all civil jury verdicts reached in the state courts of general jurisdiction in 15 jurisdictions, Moller found, among other things, that plaintiffs win slightly more than half their cases. However, the win rate ranged from a high of 66% in automobile personal injury and business cases to lows of 44% in products liability cases, and 33% in medical malpractice cases.[52]

Case outcomes often vary from jurisdiction to jurisdiction, and from data set to data set. Using a national data base on civil filings obtained by the Administrative Office of the Courts, Clermont and Eisenberg (1992) found win rates of close to 50% for products

[50] For a discussion of the possible influence of judicial ideology on case outcomes, see Ashenfelter et al. (1995).

[51] See Heise (2000), Kakalik et al. (1997).

[52] See also Peterson and Priest (1982), Shanley and Peterson (1983), and Hubbard (1987).

liability and medical malpractice cases. Surprisingly, however, they found win rates to be substantially higher before judges than before juries, despite the fact that plaintiffs select judge trials in only about 10% of the cases. Whether all of this can be explained by selection effects arising when cases vary substantially on many dimensions, remains an open question.

There is a long-line of literature evaluating the relatively abilities of juries and judges to perform their evaluative functions. Vidmar (1998) provides a useful overview of the jury literature.[53] He notes that in 1993 that only 1.8% of all civil cases filed in U.S. District Courts involved jury trial, and the fraction is very similar (2.0%) for state cases. He reviews and evaluates issues such as the ability of the jury to evaluate expert testimony, to comprehend complex legal arguments, and to understand jury instructions. With respect to the question of whether juries have a pro-plaintiff bias, the literature reaches a negative conclusion.

A recent subject of intensive study and debate has been the ability of juries to evaluate claims for punitive damages. It is well known that the distribution of such awards has a very high variance. What is less settled is the question of whether those awards are related systematically to factors raised by the issues in the case, or whether they are largely due to the particularities of juries. In one study, Hastie et al. (1998) used responses to hypothetical questions about the appropriateness of punitive damages by experimentally-chosen jurors. Vidmar criticizes this study for, among other things, focusing on the wrong questions, but is somewhat more sympathetic to Hastie and Viscusi (1998), who conclude in from a related study that juries are less able than judges to manage punitive damage deliberations.[54]

The influence or lack of influence of judges on trial outcomes has itself been a subject of some intensive empirical work.[55] In evaluating more than 200 district court cases involving the constitutionality of the U.S. Sentencing Commission in 1988, Cohen (1991) concludes that at the margin judges are likely to promote their own self-interest. Using data extracted from Jury Verdict Research's Personal Injury Verdicts and Settlements (59,304 trials and 27,429 settled cases), Helland and Tabarrok (2000) report than much of the difference in win rates (higher with judges) and awards (higher with juries) between juries and judges can be explained away by case selection effects.[56]

5.7. *Appeal*

Relatively little empirical work has concentrated on the appellate process. One important exception is Kessler et al. (1996). The authors take on the question of whether and

[53] Important if only for its historical significance is the class study of juries and judges by Kalven and Zeisel (1971).

[54] See also Viscusi (2001).

[55] For one interesting study relating to the D.C. Circuit and environmental appeals, see Revesz (2001).

[56] In a related paper, Tabarrok and Helland (1999) show that plaintiffs are more likely to win tort awards and that the awards are likely to be higher in states in which state judges are elected in partisan races than in other states (non-partisan elected judges or appointed judges).

to what extent one should expect deviations from a 50% win rate with respect to appeals, emphasizing that there are a range of factors that empirically explain deviations from the 50% result. Analyzing 3,529 appeals decided by the Seventh Circuit Court of Appeals between 1982 and 1987, they find several case characteristics that influence win rates in the manner than economics would predict. They include: differential stakes (the party that has the most at stake is more likely to win), differential information (the better informed the party, the higher the win rate), legal standard favoring one side (obvious), and agency effects (the party represented by an hourly fee lawyer will have a lower win rate at trial because the attorney will have a higher probability of taking the case to trial).

Building on our earlier discussion of differential stakes, there are reasons to expect win rates substantially below 50%. It seems clear that the probability of success on appeal is likely to be quite low for the entire population of tried cases. If the stakes (and risk aversion) were the same for both parties, then one would predict a 50% success rate on appeal. However, it seems reasonable to expect that the party that pays the cost of an appeal does so because the potential reward is quite high and the cost of an appeal is low (at least in relation to the cost of the trial itself). The implication is, given sample selection, a win rate of substantially less than 50%. Indeed, that is what Clermont and Eisenberg (2002) find in a recent analysis using the Administrative Office of the Courts database described previously. Defendants are successful in reversing loss jury trial decisions in 31% of their appeals, while losing plaintiffs success in only 13% of their appeals from jury trials. The authors point out that these results cannot be fully explained by sample selection, appealing instead to the argument that appellate judge's tend to be more favorably disposed to defendants than are either trial judges or juries.[57]

5.8. Alternatives to litigation: arbitration and mediation

Throughout the years, a wide variety of litigation reforms have been proposed and in some cases promulgated to streamline the litigation process. We have previously described a number of studies of the relatively recent attempts at the reform of the discovery process. In this subsection, we briefly highlight some of the empirical work that has focused on alternative dispute resolution systems, including arbitration and mediation.

Many studies of both arbitration and mediation have focused on the effect of various interventions on the perceived fairness of the process, rather than the economic effects on settlement behavior. Thus, in one recent study of court-connected civil case mediation programs in nine Ohio state courts, Wissler (2002) did not attempt to look at whether settlement rates themselves increased as a result of the mediation (45% of mediated cases did settle).[58] She did, however, find that few cases went to trial whether

[57] The differences are less substantial with respect to appeals of decisions by judges (23% success for defendants and 18% success for plaintiffs).

[58] Wissler (2002), p. 666. See also Kakalik et al. (1996) and Bergman and Bickeman (1998).

there was mediation or not—3% of mediated cases and 2% of non-mediated cases were actually tried. In addition, 2% of the mediated cases and 4% of the non-mediated cases were resolved by arbitration.[59] Interestingly, she also found that when mediators recommended settlement, cases were more likely to settle, while at the same time the parties were more likely to view the process as unfair. In sum, one cannot reject the possibility that successful mediated cases were those in which the parties' expectations were least divergent; it follows that successful mediated cases were those most likely to settle without mediation.

Another study focused on a four year study in the Northern District of California of an alternative dispute resolution process called early neutral evaluation. In the "experiment" half of the cases of certain types were assigned to a mediator and half were not. Rosenberg and Folberg (1994) report substantial satisfaction with the process. Moreover, in about half the early neutral evaluation cases parties reporting savings in litigation costs and shorter disposition times.

The mediation studies just described all involved interventions by the courts. An alternative form of dispute resolution is private in nature, with the parties agreeing to pay a fee to a third-party to assist with the resolution of the dispute. A useful study of these forms of mediation was offered by Rolph et al. (1994). The authors studied a wide variety of civil cases that were active in Los Angeles, California in 1992 and 1993, relying on interviews with parties and mediators, as well as lawyers' case records and Court records. They find that arbitration rather than mediation is the most popular form of alternative dispute resolution ("ADR"). While they find little or no overall effect on the civil caseload of the Los Angeles courts, they do express optimism that private forms of ADR can be successful in the future (91 former judges were offered private dispute resolution services at the time of the study).[60]

With respect to arbitration, it is typically thought that such programs can reduce delay and costs by providing a more efficient substitute for trial. But, as MacCoun (1991) has pointed out, because most disputes are resolved without trial, such programs are likely to divert more cases from settlement than from trial. The implication is that arbitration (and for that matter mediation as well) may actually increase delay and congestion in the courts. MacCoun found such a result in a study of court-annexed automobile arbitration in New Jersey. After the introduction of the arbitration program there was no reduction in the rate of cases going to trial, and an increase in the time of disposition of cases. He concludes for arbitration, as others do for mediation, that alternative dispute resolution programs appear to offer improvements in the fairness of the process, but do not necessarily increase economic efficiency, and in some cases, may make it worse.[61]

There are clearly some who believe that neither private nor public interventions of the type just described are likely to much success if any. Whether we should stick with

[59] p. 669.

[60] p. 58.

[61] For an earlier overview of the landscape, see Ebener and Betancourt (1985).

the status quo or make a more radical move remains a subject of open debate. One who argues for the latter is Hadfield (2001), who suggests that the key to achieving economic efficiency is to privatize commercial law entirely.

5.9. Fee-shifting

Under the "American rule" each party pays its own costs of litigation, whereas under the "British rule" the loser pays all. The theoretical literature on fee shifting leaves an indeterminate prediction as to the effect of a change from the American to the British rule on the probability of settlement.[62] Cooter and Rubinfeld (1989) review that literature, suggesting that in cases on the margin between trial and settlement where plaintiffs are likely to be more optimistic about expected trial outcomes, a switch to the British rule will reduce the bargaining surplus to be gained by settling, which will increase the probability of trial. However, two forces go in the opposite direction. First, under the British rule the stakes of the case are effectively increased.[63] To the extent that this encourages the parties to expend more effort at trial, the former effect will be muted or eliminated. Second, under the British rule the parties face greater risk, which would be expected to increase settlements.

Because the direction of the overall effect on the frequency of trials cannot be determined from theory alone, the empirical evidence will be decisive. The view among many lawyers that fewer suits will occur when the loser pays more is supported in part by evidence from experimental economics. In one study, Coursey and Stanley (1988) found a higher settlement rate when their experimental subjects decided disputes under the British rule as opposed to the American rule. However, in a more recent study, Main and Park (2000) found a different result—the British and American rules produced no difference in the frequency of settlements overall. As theory would predict, they did find that the British rule produces higher settlements in cases in which the plaintiff's probability of success is the highest.

There is very little actual empirical and econometric evidence concerning this topic. One interesting exception is Hughes and Snyder (1995).[64] The authors use data from a natural experiment that arose when the state of Florida switched from the American rule to the British rule for medical malpractice litigation between July 1980 and September 1985. They draw the conclusion that the British rule does tend to increase the fraction of cases settled (and the fraction that are dropped). Cases that do go to trial tend to generate more pro-plaintiff outcomes.

[62] A switch to the British system is typically thought to reduce the incentive to bring a suit; see, for example, Bebchuk and Chang (1996), Polinsky and Rubinfeld (1996).

[63] Katz (1988) simulates a 125% increase in costs.

[64] See also Snyder and Hughes (1990).

5.10. Class actions

Class actions allow an attorney or attorneys to bring a civil lawsuit on behalf of a large number of plaintiffs. Seen from the plaintiffs' perspective, class actions allow plaintiffs to recover for harms suffered when it would be uneconomic to bring individual lawsuits. From the defense perspective, class actions are seen as overdeterring defendants to the benefit of class attorneys more than the members of the class. Broader policy issues surround the overall costs of class action litigation, but also the nature of settlements (e.g., should coupon settlements be allowed?) and the extent of legal fees (how actively should courts monitor fees?).

Most empirical studies of class actions have been descriptive in nature, focusing primarily on mass tort cases. Hensler (2001) offers a recent, useful review of the literature, highlighting the results of a study of class actions conducted by the RAND Corporation's Institute for Civil Justice.[65] She points to the fact that tort cases actually represent somewhere between 9% and 18% of all class actions active in the 1995–96 period. Based on interviews with class action attorneys and documentation of active cases she offers a nuanced view that supports neither of the two extreme positions described previously. While rich in detail, the study leaves much open to future empirical work, since it does not attempt to make an overall policy assessment of the benefits and costs of the class action approach to litigation. Thus, the study found that average compensation collected or projected to be collected ranged from $270,000 to $840 million, with average payments per class member ranging from $6 to $1500 in consumer suits and $6400 to $100,000 in mass tort suits. Whether those "benefits" are worth the costs of litigation remains an unanswered question.

With respect to settlement behavior, it is believed agency issues make it likely that some cases will settle when it is in the interests of class members to go to trial (because the attorney can obtain a relatively high fee by settling and avoid the costs of trial). One early study, Rosenfield (1976), provides some supporting evidence based on a sample of over 100 class actions. More recent studies have not contradicted this view, but settlement has not typically been a focal point of the work. Thus, Willging's 1996 study for the Federal Judicial Center found that a majority of certified class actions were settled, and this rate was 2 to 5 times as high as cases that contained class actions that were not settled. However, Willging did not attempt to control for case characteristics that might distinguish the value of the underlying cases.

Acknowledgements

The authors wish to thank Ashley Ensign for her helpful research assistance, and would also like to thank Eric Rasmusen and the editors for their thoughtful comments.

[65] See Hensler et al. (1999), and Willging, Hooper, and Niemic (1996).

References

Contracts

Ayres, I. (1995). "Further Evidence of Discrimination in New Care Negotiations and Estimates of Its Cause". Michigan Law Review 94, 109–147.

Ayres, I., Gertner, R. (1989). "Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules". Yale Law Journal 99, 87–130.

Bernstein, L. (1992). "Opting Out of the Legal System: Extralegal Contractual Relations in the Diamond Industry". Journal of Legal Studies 21, 115–157.

Bernstein, L. (1996). "Merchant Law in Merchant Court: Rethinking the Code's Search for Immanent Business Norms". University of Pennsylvania Law Review 144, 1765–1822.

Bigsten, A., Collier, P., Dercon, S., Fafchamps, M., Gauthier, B., Bunning, J.W., Oduro, A., Oostendorp, R., Patillo, C., Soderbom, M., Teal, F., Zeufack, A. (2000). "Contract Flexibility and Dispute Resolution in African Manufacturing". Journal of Development Studies 36, 1–17.

Djankov, S., La Porta, R., Lopez-De-Silanes, F., Shleifer, A. (2003). "Courts". The Quarterly Journal of Economics 118, 453–517.

Farber, D.A., Matheson, J. (1985). "Beyond Promissory Estoppel: Contract Law and the Invisible Handshake". University of Chicago Law Review 52, 903–947.

Galanter, M. (2001). "Contract in Court; or Almost Everything You May or May Not Want To Know About Contract Litigation". Wisconsin Law Review 2001, 577–627.

Hillman, R.A. (1998). "Questioning the New Consensus on Promissory Estoppel: An Empirical and Theoretical Study". Columbia Law Review 98, 580–619.

Johnson, S., McMillan, J., Woodruff, C. (2002). "Courts and Relational Contracts". Journal of Law, Economics and Organization 18, 221–277.

Johnston, J.S. (1996). "The Statute of Frauds and Business Norms: A Testable Game-theoretic Model". University of Pennsylvania Law Review 144, 1859–1912.

Kahan, M., Klausner, M. (1997). "Standardization and Innovation in Corporate Contracting (or "The Economics of the Boilerplate")". Virginia Law Review 83, 713–770.

Keating, D. (2000). "Exploring the Battle of the Forms in Action". Michigan Law Review 98, 2678–2715.

Koford, K., Miller, J. (2006). "Contract Enforcement in the Early Transition of an Unstable Economy". Economic Systems 30 (1), 1–23.

Korobkin, R. (1998). "The Status Quo Bias and Contract Default Rules". Cornell Law Review 83, 608–687.

Korobkin, R. (2002). "Empirical Scholarship in Contract Law: Possibilities and Pitfalls". University of Illinois Law Review 2002, 1033–1066.

Kranton, R., Swamy, A. (1999). "The Hazards of Piecemeal Reform: British Civil Courts and the Credit Market in Colonial India". Journal of Development Economics 58, 1–24.

Lorenz, E. (1999). "Trust, Contract and Economic Cooperation". Cambridge Journal of Economics 23, 301–315.

Macaulay, S. (1963). "Non-Contractual Relations in Business: A Preliminary Study". American Sociological Review 28, 55–75.

Mann, R.J. (1996). "The First Shall be the Last: A Contextual Argument for Abandoning Temporal Rules of Lien Property". Texas Law Review 75, 11–49.

McChesney, F. (1999). "Tortious Interference with Contract versus "Efficient" Breach: Theory and Empirical Evidence". Journal of Legal Studies 28, 131–186.

McMillan, J., Woodruff, C. (1999a). "Dispute Prevention without Courts in Vietnam". Journal of Law, Economics and Organization 15, 637–658.

McMillan, J., Woodruff, C. (1999b). "Interfirm Relationships and Informal Credit in Vietnam". Quarterly Journal of Economics 114, 1285–1320.

National Center for State Courts (2002). "Tort and Contract Caseloads in State Trial Courts", http://www.ncsconline.org/D_Research/csp/1999-2000_Files/1999-2000_Tort-Contract_Section.pdf.

Pham, P.N. (1994). "The Waning of Promissory Estoppel". Cornell Law Review 79, 1263–1290.

Rice, W.E. (1992). "Judicial Bias, The Insurance Industry and Consumer Protection: An Empirical Analysis of State Supreme Courts' Bad Faith, Breach-of-Contract, Breach-of-Covenant-of-Good-Faith and Excess-Judgment Decisions". Catholic University Law Review 41, 325–382.

Schwab, S. (1988). "A Coasean Experiment on Contract Presumptions". Journal of Legal Studies 17, 237.

U.S. Department of Justice, Bureau of Justice Statistics (1995). "Tort Cases in Large Counties: Civil Justice Survey of State Courts, 1992", NCJ-153177.

U.S. Department of Justice, Bureau of Justice Statistics (1996). "Contract Cases in Large Counties, 1992", NCJ-156664.

U.S. Department of Justice, Bureau of Justice Statistics (2000a). "Contract Trials and Verdicts in Large Counties, 1996", NCJ-179451.

Weintraub, R.J. (1992). "A Survey of Contract Practice and Policy". Wisconsin Law Review 1992, 1–60.

Torts: Introduction and medical malpractice

Barker, D.K. (1992). "The Effects of Tort Reform on Medical Malpractice Insurance Markets: An Empirical Analysis". Journal of Health Politics, Policy and Law 17, 143–161.

Baldwin, L.M., Hart, L.G., Lloyd, M., Fordyce, M., Rosenblatt, R.A. (1995). "Defensive Medicine and Obstetrics". The Journal of the American Medical Association 274, 1606–1610.

Brennan, T.A., Sox, C.M., Burstin, H.R. (1996). "Relation Between Negligent Adverse Events and the Outcomes of Medical-Malpractice Litigation". New England Journal of Medicine 335, 1963–1967.

Craswell, R., Calfee, J. (1986). "Deterrence and Uncertain Legal Standards". Journal of Law, Economics, and Organization 2, 279–303.

Dewees, D., Duff, D., Trebilcock, M. (1996). Exploring the Domain of Accident Law: Taking the Facts Seriously. Oxford University Press, New York.

Danzon P. (1982). "The Frequency and Severity of Medical Malpractice Claims", RAND R-2870-ICJ/HCFA.

Danzon, P. (1986). "New Evidence on the Frequency and Severity of Malpractice Claims". Law and Contemporary Problems 5, 57–84.

Danzon, P. (1991). "Liability for Medical Malpractice". Journal of Economic Perspectives 5, 51–69.

Danzon, P. (2000). "Liability for Medical Malpractice". In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics, vol. 1B. Elsevier, New York, ch. 26.

Danzon, P., Lillard, L. (1983). "Settlement out of Court: The Disposition of Medical Malpractice Claims". Journal of Legal Studies 12, 345–377.

Dubay, L., Kaestner, R., Waidmann, T. (1999). "The Impact of Malpractice Fears on Cesarean Section Rates". Journal of Health Economics 18, 491–522.

Dubay, L., Kaestner, R., Waidmann, T. (2001). "Medical Malpractice Liability and Its Effect on Prenatal Care Utilization and Infant Health". Journal of Health Economics 20, 591–611.

Eaton, T.A., Talarico, S.M., Dunn, R.E. (2000). "Another Brick in the Wall: An Empirical Look at Georgia Tort Litigation in the 1990s". Georgia Law Review 34, 1056–1154.

Eisenberg, T., Schwab, S. (1987). "The Reality of Constitutional Tort Litigation". Cornell Law Review 72, 641–695.

Farber, H.S., White, M.J. (1991). "Medical Malpractice: An Empirical Examination of the Litigation Process". RAND Journal of Economics 22, 199–217.

Galanter, M. (1996). "Real World Torts: An Antidote to Anecdote". Maryland Law Review 55, 1093–1160.

Harvard Medical Practice Study (1990). "Patients, Doctors, and Lawyers: Medical Injury, Malpractice Litigation, and Patient Compensation in New York", a report by the Harvard Medical Practice Study to the State of New York. The President and Fellows of Harvard College, Cambridge, MA.

Havighurst, C.C. (1995). Health Care Choices: Private Contracts as Instruments of Health Care Reform. AEI Press, Washington, DC.

Hellinger F.J., Encinosa W.E. (2003), "The Impact of State Laws Limiting Malpractice Awards on the Geographic Distribution of Physicians", Agency for Healthcare Research and Quality. (Available at: www.ahrq.gov/research/tortcaps/tortcaps.htm#Introduction.)

Institute of Medicine (2000). To Err is Human: Building a Safer Health System. National Academies Press, Washington, DC.

Kessler, D.P., McClellan, M.B. (1996). "Do Doctors Practice Defensive Medicine?". Quarterly Journal of Economics 111, 353–390.

Kessler, D.P., McClellan, M.B. (1997). "The Effects of Malpractice Pressure and Liability Reforms on Physicians' Perceptions of Medical Care". Law and Contemporary Problems 60, 81–106.

Kessler, D.P., McClellan, M.B. (2002a). "How Liability Reform Affects Medical Productivity". Journal of Health Economics 21, 931–955.

Kessler, D.P., McClellan, M.B. (2002b). "Malpractice Law and Health Care Reform: Optimal Liability Policy in an Era of Managed Care". Journal of Public Economics 84, 175–197.

Klingman, D., Localio, A.R., Sugarman, J., Wagner, J.L., Polishuk, P.T., Wolfe, L., Corrigan, J.A. (1996). "Measuring Defensive Medicine Using Clinical Scenario Surveys". Journal of Health Politics, Policy and Law 21 (2), 185–217.

Localio, A.R., Lawthers, A.G., Bengtson, J.M., Hebert, L.E., Weaver, S.L., Brennan, T.A., Landis, J.R. (1993). "Relationship between Malpractice Claims and Cesarean Delivery". The Journal of the American Medical Association 269 (3), 366–373.

Moser, J., Musaccio, R. (1991). "The Cost of Medical Professional Liability in the 1980s". Medical Practice Management, 3–9.

Office of Technology Assessment, U.S. Congress (1993). Impact of Legal Reforms on Medical Malpractice Costs, OTA-BP-H-119. U.S. Government Printing Office, Washington, DC.

Office of Technology Assessment, U.S. Congress (1994). Defensive Medicine and Medical Malpractice, OTA-H-602. U.S. Government Printing Office, Washington, DC.

Priest, G.L. (1988). "Understanding the Liability Crisis, in New Directions Liability Law". In: Proceedings of the Academy of Political Science 37(1): New Directions in Liability Law. Academy of Political Science, New York, NY, pp. 196–211.

Reynolds, R.A., Rizzo, J.A., Gonzales, M.L. (1987). "The Cost of Medical Professional Liability". The Journal of the American Medical Association 257, 2776–2781.

Rock, S.M. (1988). "Malpractice Premiums and Primary Cesarean Section Rates in New York and Illinois". Public Health Reporter 58, 459–468.

Schwartz, G.T. (1981). "The Vitality of Negligence and the Ethics of Strict Liability". Georgia Law Review 15, 963–1004.

Schwartz, G.T. (1992). "The Beginning and the Possible End of the Rise of Modern American Tort Law". Georgia Law Review 26, 601–702.

Shanley, M., Peterson, M.A. (1987). Posttrial Adjustments to Jury Awards. RAND Corporation, Santa Monica, CA.

Shavell, S. (1987). The Economic Analysis of Accident Law. Harvard University Press, Cambridge, MA.

Sloan, F.A. (1990). "Experience Rating: Does It Make Sense for Medical Malpractice Insurance?". American Economic Review 80, 128–133.

Sloan, F.A., Mergenhagen, P.M., Bovbjerg, R.R. (1989). "Effects of Tort Reforms on the Value of Closed Medical Malpractice Claims: A Microanalysis". Journal of Health Politics, Policy and Law 14, 663–689.

Studdert, D.M., Brennan, T. (2001). "Toward a Workable Model of 'No-fault' Compensation for Medical Injury in the United States". American Journal of Law and Medicine 27, 225–252.

Studdert, D.M., Brennan, T., Thomas, E.J. (2000). "Beyond Dead Reckoning: Measures of Medical Injury Burden, Malpractice Litigation, and Alternative Compensation Models from Utah and Colorado". Indiana Law Review 33, 1643–1686.

Studdert, D.M., Thomas, E.J., Zbar, B.I., Newhouse, J.P., Weiler, P.C., Brennan, T.A. (1997). "Can the United States Afford a 'No-fault' System of Compensation for Medical Injury?". Law and Contemporary Problems 60 (2), 1–34.

U.S. Department of Justice, Bureau of Justice Statistics (1995). "Tort Cases in Large Counties", NCJ-153177.

U.S. Department of Justice, Bureau of Justice Statistics (1997). "Federal Tort Trials and Verdicts, 1994–95", NCJ-165810.

U.S. Department of Justice, Bureau of Justice Statistics (1999). "Federal Tort Trials and Verdicts, 1996–97", NCJ-172855.

U.S. Department of Justice, Bureau of Justice Statistics (2000b). "Tort Trials and Verdicts in Large Counties, 1996", NCJ-179769.

Weiler, P.C., Hiatt, H.H., Newhouse, J.P., Johnson, W.G., Brennan, T.A., Leape, L.L. (1993). A Measure of Malpractice: Medical Injury, Malpractice Litigation, and Patient Compensation. Harvard University Press, Cambridge, MA.

Zuckerman, S., Bovbjerg, R.R., Sloan, F. (1990). "Effects of Tort Reforms and Other Factors on Medical Malpractice Insurance Premiums". Inquiry 27, 167–182.

Torts: Automobile bodily injury

Abrahamse, A.F., Carroll, S.J. (1995). The Effects of a Choice Auto Insurance Plan on Insurance Costs. RAND Institute for Civil Justice, Santa Monica, CA.

Carroll, S.J., Abrahamse, A. (1999). The Effects of a Choice Automobile Insurance Plan on Insurance Costs and Compensation: An Analysis Based on 1997 Data. RAND Institute for Civil Justice, Santa Monica, CA.

Carroll, S.J., Kakalik, J.S., Pace, N.M., Adams, J.L. (1991). No-Fault Approaches to Compensating People Injured in Automobile Accidents. RAND Institute for Civil Justice, Santa Monica, CA.

Chaloupka, F.J., Saffer, H., Grossman, M. (1993). "Alcohol-Control Policies and Motor-Vehicle Fatalities". Journal of Legal Studies 22, 161–186.

Cohen A., Dehejia R. (2003). "The Effect of Automobile Insurance and Accident Liability Laws on Traffic Fatalities", NBER Working Paper 9602.

Cummins, J.D., Phillips, R.D., Weiss, M.A. (2001). "The Incentive Effects of No-Fault Automobile Insurance". Journal of Law and Economics 44, 427–464.

Devlin, A.R. (1990). "Some Welfare Implications of No-Fault Automobile Insurance". International Review of Law and Economics 10, 193–205.

Edlin, A.S. (2003). "Per-Mile Premiums for Auto Insurance". In: Arnott, R., et al. (Eds.), Economics for an Imperfect World: Essays In Honor of Joseph Stiglitz. MIT Press, Cambridge, MA, pp. 53–82.

Kabler, B. (1999). "The Case Against Auto Choice". Journal of Insurance Regulation 18, 53–79.

Kakalik, J.S., Robyn, A.E. (1982). Costs of the Civil Justice System: Court Expenditures for Processing Tort Cases. RAND Institute for Civil Justice, Santa Monica, CA.

Kessler, D. (1995). "Fault, Settlement, and Negligence Law". RAND Journal of Economics 26, 296–313.

Kochanowski, P.S., Young, M.V. (1985). "Deterrent Aspects of No-Fault Automobile Insurance: Some Empirical Findings". Journal of Risk and Insurance 52, 269–288.

Landes, E.M. (1982). "Insurance, Liability, and Accidents: A Theoretical and Empirical Investigation of the Effect of No-Fault on Accidents". Journal of Law and Economics 25, 49–65.

Loughran, D.S. (2001). The Effect on No-Fault Automobile Insurance on Driver Behavior and Automobile Accidents in the United States. RAND Institute for Civil Justice, Santa Monica, CA.

McEwin, I. (1989). "No-Fault and Road Accidents: Some Australian Evidence". International Review of Law and Economics 9, 13–24.

O'Connell, J., Carroll, S., Horowitz, M., Abrahamse, A. (1993). "Consumer Choice in the Auto Insurance Market". Maryland Law Review 52, 1016–1062.

O'Connell, J., Carroll, S., Horowitz, M., Abrahamse, A., Jamieson, P. (1996). "The Comparative Costs of Allowing Consumer Choice for Auto Insurance in all Fifty States". Maryland Law Review 55, 160–222.

O'Connell, J., Carroll, S., Horowitz, M., Abrahamse, A., Kaiser, D. (1995). "The Costs of Consumer Choice For Auto Insurance in States Without No-Fault Insurance". Maryland Law Review 54, 281–351.

O'Connell, J., Joost, R. (1986). "Giving Motorists a Choice Between Fault and No-Fault Insurance". Virginia Law Review 72, 61–90.

Rolph, J.E., Hammit, J.K., Houchens, R.L., Polin, S.S. (1985). Automobile Accident Compensation. Who Pays How Much How Soon, vol. I. RAND Institute for Civil Justice, Santa Monica, CA.

Ruhm, C.J. (1996). "Alcohol Policies and Highway Vehicle Fatalities". Journal of Health Economics 15, 435–454.

Sloan, F.A., Reilly, B.A., Schenzler, C. (1994). "Effects of Prices, Civil and Criminal Sanctions, and Law Enforcement on Alcohol-Related Mortality". Journal of Studies on Alcohol 55, 454–465.

Sloan, F.A., Reilly, B.A., Schenzler, C. (1995). "Effects of Tort Liability and Insurance on Heavy Drinking and Drinking and Driving". Journal of Law and Economics 38, 49–77.

Urban Institute (1991). "The Costs of Highway Crashes: Final Report, June". Urban Institue Press, Washington, DC.

U.S. Department of Transportation (2003). "Traffic Safety Facts 2002", December.

Vickrey, W. (1968). "Automobile Accidents, Tort Law, Externalities, and Insurance: An Economist's Critique". Law and Contemporary Problems 33, 464.

White, M. (1989). "An Empirical Test of the Comparative and Contributory Negligence Rules in Accident Law". RAND Journal of Economics 20, 308–330.

Wittman, D. (1986). "The Price of Negligence Under Differing Liability Rules". Journal of Law and Economics 29, 151–163.

Zador, P., Lund, A. (1986). "Re-Analyses of the Effects of No-Fault Auto Insurance on Fatal Crashes". The Journal of Risk and Insurance 53, 226–241.

Torts: Products liability

Campbell, T.J., Kessler, D.P., Shepherd, G.B. (1998). "The Link between Liability Reforms and Productivity: Some Empirical Evidence". In: Brookings Papers on Economic Activity, Microeconomics. The Brookings Institution, Washington, D.C., pp. 107–148.

Congressional Budget Office, The Congress of the United States (2003). "The Economics of U.S. Tort Liability: A Primer", October.

Eads, G., Reuter, P. (1985). Designing Safer Products: Corporate Responses to Product Liability Law and Regulation. RAND Institute for Civil Justice, Santa Monica, CA.

Eisenberg, T. (1999). "Judicial Decision-Making in Federal Products Liability Cases, 1978–1997". DePaul Law Review 49, 323–333.

Eisenberg, T., Henderson, J.A. Jr. (1992). "Inside the Quiet Revolution in Products Liability". UCLA Law Review 39, 731–810.

Eisenberg, T., Goerdt, J., Ostrom, B., Rottman, D. (1996). "Federal Product Liability Litigation Reform: Recent Developments and Statistics, Litigation Outcomes in State and Federal Courts: A Statistical Portrait". Seattle University Law Review 19, 433–453.

Garber, S. (1993). Product Liability and the Economics of Pharmaceuticals and Medical Devices. RAND Institute for Civil Justice, Santa Monica, CA.

Henderson, J.A. Jr. (1991). "Judicial Reliance on Public Policy: An Empirical Analysis of Products Liability Decisions". George Washington Law Review 59, 1570–1613.

Henderson, J.A. Jr., Eisenberg, T. (1990). "The Quiet Revolution in Products Liability: An Empirical Study of Legal Change". UCLA Law Review 37, 479–553.

Johnson, R.B. (1991). "The Impact of Liability on Innovation in the Chemical Industry". In: Huber, P.W., Litan, R.E. (Eds.), The Liability Maze: The Impact of Liability Law on Safety and Innovation. The Brookings Institution, Washington, D.C., pp. 428–455.

Keeton, W.P., Dobbs, D.B., Keeton, R.E., Owen, D.G. (1984). Prosser and Keeton on Torts, 5th edn. West Publishing Co., St. Paul, MN.

Litan R.E. (1991). "The Safety and Innovation Effects of U.S. Liability Law: The Evidence", AEA Papers and Proceedings, May, pp. 59–64.

Mackay, M. (1991). “Liability, Safety, and Innovation in the Automotive Industry”. In: Huber, P.W., Litan, R.E. (Eds.), The Liability Maze: The Impact of Liability Law on Safety and Innovation. The Brookings Institution, Washington, D.C., pp. 191–223.

Manning, R.L. (1994). “Changing Rules in Tort Law and the Market for Childhood Vaccines”. Journal of Law and Economics 37, 247–275.

Manning, R.L. (1997). “Products Liability and Prescription Drug Prices in Canada and the United States”. Journal of Law and Economics 40, 203–240.

Martin, R. (1991). “General Aviation Manufacturing: An Industry under Siege”. In: Huber, P.W., Litan, R.E. (Eds.), The Liability Maze: The Impact of Liability Law on Safety and Innovation. The Brookings Institution, Washington, D.C., pp. 478–499.

Priest, G.L. (1988). “Products Liability Law and the Accident Rate”. In: Litan, R.E., Winston, C. (Eds.), Liability: Perspectives and Policy. The Brookings Institution, Washington, D.C., pp. 184–222.

Viscusi, W.K. (1990). “The Performance of Liability Insurance in States with Different Products-Liability Statutes”. Journal of Legal Studies 19, 809–836.

Viscusi, W.K. (1991). “The Dimensions of the Product Liability Crisis”. Journal of Legal Studies 20, 147–177.

Viscusi, W.K., Moore, M.J. (1993). “Product Liability, Research and Development, and Innovation”. Journal of Political Economics 101, 161–184.

Torts: Mass torts

Carroll, S.J., Hensler, D., Abrahamse, A., Gross, J., White, M., Ashwood, S., Sloss, E. (2002). Asbestos Litigation Costs and Compensation: An Interim Report. RAND Institute for Civil Justice, Santa Monica, CA.

Hensler, D. (2001). “Multi-district Litigation and Aggregation Alternatives: The Role of Muli-Districting in Mass Tort Litigation: An Empirical Investigation”. Seton Hall Law Review 31, 883–906.

Hensler, D., Peterson, M.A. (1993). “Understanding Mass Personal Injury Litigation: A Socio-Legal Analysis”. Brooklyn Law Review 59, 961.

Hensler, D., Felstiner, W.L., Selvin, M., Ebener, P.A. (1985). Asbestos in the Courts: The Challenge of Mass Toxic Torts. RAND Institute for Civil Justice, Santa Monica, CA.

Peterson, M., Selvin, M. (1988). “Resolution of Mass Torts: Toward a Framework for Evaluation of Aggregative Procedures”. N-2805-ICJ. RAND Institute for Civil Justice, Santa Monica, CA.

Torts: Punitive damages

Baker, T. (1998). “Transforming Punishment into Compensation: In the Shadow of Punitive Damages”. Wisconsin Law Review 1998, 211–236.

Baron, J., Ritov, I. (1993). “Intuitions about Penalties and Compensation in the Context of Tort Law”. Journal Risk and Uncertainty 7, 17.

Daniels, S., Martin, J. (1990). “Myth and Reality in Punitive Damages”. Minnesota Law Review 75, 1–64.

Eisenberg, T. (1998). “Responses: Measuring the Deterrent Effect of Punitive Damages”. Georgetown Law Journal 87, 347–360.

Eisenberg, T., Goerdt, J., Ostrom, B., Rottman, D., Wells, M.T. (1997). “The Predictability of Punitive Damages”. Journal of Legal Studies 26, 623–661.

Eisenberg, T., LaFountain, N., Ostrom, B., Rottman, D., Wells, M.T. (2002). “Juries, Judges, and Punitive Damages: An Empirical Study”. Cornell Law Review 87, 743–782.

GAO Report (1989). “Product Liability: Verdicts and Case Resolution in Five States”, United States General Accounting Office, September.

Hensler, D., Moller, E. (1995). Trends in Punitive Damages: Preliminary Data from Cook County, Illinois and San Francisco, California. RAND Institute for Civil Justice, Santa Monica, CA.

Insurance Services Office (1988). “Claim File Data Analysis: Technical Analysis of Study Results”, December.

Karpoff, J.M., Lott, J.R. (1999). “On the Determinants and Importance of Punitive Damage Awards”. Journal of Law and Economics 42, 527–572.

Koenig, T. (1998). “The Shadow Effect of Punitive Damages on Settlements”. Wisconsin Law Review 1998, 169–209.

Launie, J.J., Jennings, W.P., Witt, R.C. (1994). The Economic Impact of Punitive Damages in Texas: Carpet-Bombing the State’s Prosperity. Texas Public Policy Foundation, Austin, TX.

Moller, E., Pace, N.M., Carroll, S.J. (1997). Punitive Damages in Financial Injury Verdicts: An Executive Summary. RAND Institute for Civil Justice, Santa Monica, CA.

Peterson, M., Sarma, S., Shanley, M. (1987). Punitive Damages: Empirical Findings. RAND Institute for Civil Justice, Santa Monica, CA.

Polinsky, M. (1997). “Are Punitive Damages Really Insignificant, Predictable, and Rational? A Comment on Eisenberg et al.”. Journal of Legal Studies 26, 663–677.

Polinsky, M., Shavell, S. (1998). “Punitive Damages: An Economic Analysis”. Harvard Law Review 111, 869–962.

Robbennolt, J.K. (2002). “Determining Punitive Damages: Empirical Insights and Implications for Reform”. Buffalo Law Review 50, 103–203.

Sunstein, C.R., Kahneman, D., Schkade, D. (1998). “Assessing Punitive Damages (With Notes on Congnition and Valuation in Law)”. Yale Law Journal 107, 2071–2153.

Viscusi, W.K. (1998). “Punitive Damages: The Social Costs of Punitive Damages Against Corporations in Environmental and Safety Torts”. Georgetown Law Journal 87, 285–346.

Viscusi, W.K. (2001a). “Jurors, Judges, and the Mistreatment of Risk by the Courts”. Journal of Legal Studies 30, 107–142.

Viscusi, W.K. (2001b). “The Challenge of Punitive Damages Mathematics”. Journal of Legal Studies 30, 313–350.

Viscusi, W.K. (2002). “Corporate Risk Analysis: A Reckless Act?”. Stanford Law Review 52, 547–597.

Property

Agnello, R.J., Donnelley, L.P. (1975). “Property Rights and Efficiency in the Oyster Industry”. Journal of Law and Economics 18, 521–533.

Alston, L.J., Libecap, G.D., Schneider, R. (1996). “The Determinants and Impact of Property Rights: Land Titles on the Brazilian Frontier”. Journal of Law, Economics and Organizations 12 (1), 25–61.

Anderson, C.L., Swimmer, E. (1997). “Some Empirical Evidence on Property Rights of First Peoples”. Journal of Economic Behavior & Organization 33, 1–22.

Anderson, T.L., Hill, P.J. (1990). “The Race for Property Rights”. Journal of Law and Economics 33 (April), 177–197.

Anderson, T.L., Lueck, D. (1992). “Land Tenure and Agricultural Productivity on Indian Reservations”. Journal of Law and Economics 35 (2), 427–454.

Bailey, M.J. (1992). “Approximate Optimality of Aboriginal Property Rights”. Journal of Law and Economics 35 (April), 183–198.

Barro, R.J. (1991). “Economic Growth in a Cross Section of Countries”. Quarterly Journal of Economics 106, 407–443.

Calabresi, G., Melamed, D.A. (1972). “Property Rules, Liability Rules, and Inalienability: One View of the Cathedral”. Harvard Law Review 85, 1089–1128.

Coase, R. (1960). “The Problem of Social Cost”. Journal of Law and Economics 3, 1–44.

De Alessi, M. (1998). Fishing for Solutions. Coronet Books, Philadelphia.

Demsetz, H. (1967). “Toward a Theory of Property Rights”. American Economic Review 57 (2), 347–359.

Eggertsson, T. (1992). “Analyzing Institutional Successes and Failures: A Millennium of Common Mountain Pastures in Iceland”. International Review of Law and Economics 12 (4), 423–437.

Ellickson, R.C. (1991). Order without Law. Harvard University Press, Cambridge.

Ellickson, R.C. (1993). “Property in Land”. Yale Law Journal 102 (4), 1315–1400.

Epstein, R.A. (2001). The Allocation of the Commons: Parking and Stopping on the Commons, Olin Working Paper No. 134/Public Law Working Paper No. 15. University of Chicago Law & Economics Program, Chicago.

Farnsworth, W. (1999). "Do Parties to Nuisance Cases Bargain After Judgment? A Glimpse Inside the Cathedral". University of Chicago Law Review 66, 373–426.

Fischel, W.A. (1990). Do Growth Controls Matter? A Review of Empirical Evidence on the Effectiveness and Efficiency of Local Government Land Use Regulation. Lincoln Institute of Land Policy, Cambridge.

Grafton, Q., Squires, D., Fox, K.J. (2000). "Private Property and Economic Efficiency: A Study of Common-Pool Resource". Journal of Law and Economics 43 (Oct), 679–713.

Grafton, Q., Squires, D., Kirkey, J.E. (1996). "Turning the Tide? Private Property Rights and the Crisis in World Fisheries". Contemporary Economic Policy 14, 90–99.

Heller, M.A. (1998). "The Tragedy of the Anticommons: Property in the Transition from Marx to Markets". Harvard Law Review 111 (Jan), 621–688.

Joskow, P.L., Schmalensee, R. (1998). "The Political Economy of Market-Based Environmental Policy: The U.S. Acid Rain Program". Journal of Law and Economics 41, 37–83.

Joskow, P.L., Schmalensee, R., Bailey, E.M. (1998). "The Market for Sulfur Dioxide Emissions". American Economic Review 88, 669–685.

Kanazawa, M.T. (1998). "Efficiency in Western Water Law: The Development of the California Doctrine, 1850–1911". Journal of Legal Studies 27 (1), 159–185.

Keefer, P., Knack, S. (1997). "Why Don't Poor Countries Catch Up? A Cross-National Test of an Institutional Explanation". Economic Inquiry 35, 590–602.

Knack, S., Keefer, P. (1995). "Institutions and Economic Performance: Cross-Country Tests Using Alternative Institutional Measures". Economics and Politics 7, 207–227.

Lanjouw, J.O., Levy, P.I. (2002). "Untitled: A Study of Formal and Informal Property Rights in Urban Ecuador". Economic Journal 112 (482), 986–1019.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R. (1999). "The Quality of Government". Journal of Law, Economics, and Organizations 15 (1), 222–279.

Libecap, G.D. (1989). "Distributional Issues in Contracting for Property Rights". Journal of Institutional and Theoretical Economics 145, 6–24.

Libecap, G.D., Wiggins, S.N. (1984). "Contractual Responses to the Common Pool: Prorationing of Crude Oil Production". American Economic Review 74 (1), 87–98.

Lueck, D. (1994). "Common Property as an Egalitarian Share Contract". Journal of Economic Behavior and Organizations 25, 93–108.

Lueck, D. (1995). "The Rule of First Possession and the Design of the Law". Journal of Law and Economics 38, 393–436.

Miceli, T.J., Sirmans, C.F. (2000). "Partitial of Real Estate; or, Breaking up is (not) Hard to Do". Journal of Legal Studies 29 (2), 783–796.

Miceli, T.J., Sirmans, C.F., Kieyah, J. (2001). "The Demand for Land Title Registration: Theory and Evidence from Kenya". American Law and Economics Review 3 (2), 275–287.

Migot-Adholla, S., Hazell, P., Blarel, B., Place, F. (1991). "Indigenous Land Rights Systems in Sub-Saharan Africa: A Constraint on Productivity?". The World Bank Economic Review 5 (1), 155–175.

North, D.C. (1981). Structure and Change in Economic History. W.W. Norton, New York, NY.

North, D.C. (1990). Institutions, Institutional Change, and Economic Performance. Cambridge University Press, Cambridge, MA.

Norton, S.W. (1998). "Poverty, Property Rights, and Human Well-Being: A Cross National Study". Cato Journal 18 (2), 233–245.

Pirrong, S.C. (1995). "The Efficient Scope of Private Transactions-Cost-Reducing Institutions: The Successes and Failures of Commodity Exchanges". Journal of Legal Studies 24 (1), 229–255.

Rose-Ackerman, S. (1985). "Inalienability and the Theory of Property Rights". Columbia Law Review 85, 931–969.

Shleifer, A., Boycko, M., Vishny, R. (1993). "Privatizing Russia". Brookings Papers on Economic Activity 2, 139–192.

Shleifer, A., Boycko, M., Vishny, R. (1995). Privatizing Russia. MIT Press, Cambridge, 1995.

Schmalensee, R., Joskow, P.L., Denny Ellerman, A., Montero, J.P., Bailey, E.M. (1998). "An Interim Evaluation of Sulfur Dioxide Emissions Trading". Journal of Economic Perspectives 12, 53–68.

Townsend, R.E. (1990). "Entry Restrictions in the Fishery: A Survey of the Evidence". Land Economics 66, 360–378.

Troesken, W. (1997). "The Sources of Public Ownership: Historical Evidence from the Gas Industry". Journal of Law, Economics and Organizations 13 (1), 1–25.

Legal process

Alexander, J.C. (1991). "Do the Merits Matter? A Study of Settlement in Securities Class Actions". Stanford Law Review 43, 497–598.

Ashenfelter, O., Eisenberg, T., Schwab, S.J. (1995). "Politics and the Judiciary: The Influence of Judicial Background on Case Outcomes". Journal of Legal Studies 24, 257–281.

Bebchuk, L., Chang, H.F. (1996). "An Analysis of Fee Shifting Based on the Margin of Victory: On Frivolous Suits, Meritorious Suits, and the Role of Rule 11". Journal Legal Studies 25, 371–403.

Bebchuk, L., Cohen, A., Ferrell, A. (2002). "Does the Evidence Favor State Competition in Corporate Law?". California Law Review 90 (6), 1775–1822.

Bergman, E.J., Bickerman, J.G. (Eds.) (1998). Court-Annexed Mediation: Critical Perspectives on Selected State and Federal Programs. Pike & Fischer, Inc., Bethesda, MD.

Borchers, P.J. (1992). "The Choice-of-Law Revolution: An Empirical Study". Washington and Lee Law Review 49, 357–384.

Clark, D.S. (1990). "Civil Litigation Trends in Europe and Latin America since 1945: The Advantage of Intracountry Comparisons". Law and Society Review 24 (2), 549–569.

Clermont, K.M., Eisenberg, T. (1992). "Trial by Jury or Judge: Transcending Empiricism". Cornell Law Review 77, 1124–1177.

Clermont, K.M., Eisenberg, T. (1995). "Exorcising the Evil of Forum-Shopping". Cornell Law Review 80, 1507–1535.

Clermont, K.M., Eisenberg, T. (1998). "Do Case Outcomes Really Reveal Anything About the Legal System? Win Rates and Removal Jurisdiction". Cornell Law Review 83, 581–607.

Clermont, K.M., Eisenberg, T. (2002). "Litigation Realities". Cornell Law Review 88, 119–154.

Cohen, M. (1991). "Explaining Judicial Behavior, or What's 'Unconstitutional' about the Sentencing Commission". Journal of Law, Economics & Organization 7 (1), 183–199.

Connolly, P.R., Holleman, E., Kuhlman, M. (1978). Judicial Controls and the Civil Litigative Process: Discovery. Federal Judicial Center, Washington, DC.

Cooter, R.D., Rubinfeld, D.L. (1989). "Economic Analysis of Legal Disputes and Their Resolution". Journal of Economic Literature 27 (3), 1067–1097.

Cooter, R.D., Rubinfeld, D.L. (1990). "Trial Courts: An Economic Perspective". Law and Society Review 24, 533–546.

Cooter, R.D., Rubinfeld, D.L. (1994). "An Economic Model of Legal Discovery". Journal of Legal Studies 23, 435–463.

Coursey, D., Stanley, L.R. (1988). "Pre-trial Bargaining Behavior within the Shadow of the Law: Theory and Experimental Evidence". International Review of Law and Economics 8, 161–179.

Danzon, P.M., Lillard, L.A. (1983). "Settlement Out of Court: The Disposition of Medical Malpractice Claims". Journal of Legal Studies 12 (2), 345–377.

Ebener, P.A., Betancourt, D.R. (1985). Court-Annexed Arbitration: The National Picture. RAND Institute for Civil Justice, Santa Monica, CA.

Eisenberg, T. (1990). "Testing the Selection Effect: A New Theoretical Framework with Empirical Tests". Journal of Legal Studies 19, 337–358.

Eisenberg, T., Clermont, K.M. (1996). "Trial by Jury or Judge: Which is Speedier?". Judicature 79, 176–180.

Eisenberg, T., Farber, H.S. (1997). "The Litigious Plaintiff Hypothesis: Case Selection and Resolution". RAND Journal of Economics 28, 92–112.

Farber, H.S., White, M.J. (1991). “Medical Malpractice: An Empirical Examination of the Litigation Process”. RAND Journal of Economics 22, 199–217.

Fournier, G.M., Zuehlke, T.W. (1989). “Litigation and Settlement: An Empirical Approach”. Review of Economics and Statistics 71, 189–195.

Galanter, M. (1986). “The Day after the Litigation Explosion”. Maryland Law Review 46, 3–39.

Galanter, M., Cahill, M. (1994). “ ‘Most Cases Settle’: Judicial Promotion and Regulation of Settlements”. Stanford Law Review 46, 1339–1391.

Glaser, W.A. (1968). Pretrial Discovery and the Adversary System. Russell Sage Foundation, New York.

Gross, S.R., Syverud, K.D. (1991). “Getting to No: A Study of Settlement Negotiations and the Selection of Cases for Trial”. Michigan Law Review 90, 319–393.

Gross, S.R., Syverud, K.D. (1996). “Don’t Try: Civil Jury Verdicts in a System Geared to Settlement”. UCLA Law Review 44, 1–64.

Hadfield, G.K. (2001). “Privatizing Commercial Law”. Regulation Magazine 24 (1), 40–45.

Hastie, R., Viscusi, W.K. (1998). “What Juries Can’t Do Well: The Jury’s Performance as a Risk Manager”. Arizona Law Review 40, 901–921.

Hastie, R., Schkade, D., Payne, J. (1998). “A Study of Juror and Jury Judgments in Civil Cases: Deciding Liability for Punitive Damages”. Law & Human Behavior 22, 287–314.

Hay, B.L. (1994). “Civil Discovery: Its Effects and Optimal Scope”. Journal of Legal Studies 23, 259–278.

Hay, B.L., Spier, K.E. (1998). “Settlement of Litigation”. In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 3. Macmillan Reference Limited, London, pp. 442–451.

Heise, M. (2000). “Justice Delayed?: An Empirical Analysis of Civil Case Disposition Time”. Case Western Reserve Law Review 50 (4), 813–849.

Helland, E., Tabarrok, A. (2000). “Runaway Judges? Selection Effects and the Jury”. Journal of Law, Economics and Organization 16 (2), 306–333.

Hensler, D. (2001). “Revisiting the Monster: New Myths and Realities of Class Action and other Large Scale Litigation”. Duke Journal of Comparative & International Law 11, 179–213.

Hensler, D., Dombey-Moore, B., Giddens, B., Gross, J., Moller, E.K., Pace, N.M. (1999). Class Action Dilemmas: Pursuing Public Goals for Private Gains, Executive Summary. RAND Institute for Civil Justice, Santa Monica, CA.

Huang, K.-C. (2000). “Mandatory Disclosure: A Controversial Device with No Effects”. Pace Law Review 21, 203.

Hubbard, F.P. (1987). “‘Patterns’ in Civil Jury Verdicts in the State Circuit Courts of South Carolina: 1976–1985”. South Carolina Law Review 38, 699–754.

Hughes, J.W. (1989). “The Effect of Medical Malpractice Reform Laws on Claim Disposition”. International Review of Law and Economics 9, 57–78.

Hughes, J.W., Snyder, E.A. (1995). “Litigation and Settlement under the English and American Rules: Theory and Evidence”. Journal of Law and Economics 38, 225–250.

Johnston, J.S., Waldfogel, J. (2002). “Does Repeat Play Elicit Cooperation? Evidence from Federal Civil Litigation”. Journal of Legal Studies 31, 39–60.

Kakalik, J.S., Hensler, D.R., McCaffrey, D., Oshiro, M., Pace, N.M., Vaiana, M.E. (1998). Discovery Management: Further Analysis of the Civil Justice Reform Act Evaluation Data. RAND Institute for Civil Justice, Santa Monica, CA.

Kakalik, J.S., Dunworth, T., Hill, L.A., McCaffrey, D., Oshiro, M., Pace, N.M., Vaiana, M.E. (1996). An Evaluation of Mediation and Early Neutral Evaluation under the Civil Justice Reform Act. RAND Institute for Civil Justice, Santa Monica, CA.

Kakalik, J.S., Dunworth, T., Hill, L.A., McCaffrey, D., Oshiro, M., Pace, N.M., Vaiana, M.E. (1997). “Just, Speedy, and Inexpensive? An Evaluation of Judicial Case Management under the Civil Justice Reform Act”. Alabama Law Review 49 (1), 17.

Kalven, H. Jr., Zeisel, H. (1971). The American Jury, 2nd edn. University of Chicago Press, Chicago.

Kaplow, L., Shavell, S. (2002). “Economic Analysis of Law”. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. III. Elsevier Science, Amsterdam, Chapter 25.

Katz, A. (1988). "Judicial Decisionmaking and Litigation Expenditure". International Review of Law and Economics 8, 127–143.

Keilitz, S., Hanson, R.A., Daley, H.W.K. (1993). "Is Civil Discovery in State Trial Courts out of Control?". State Court Journal 17, 8.

Kessler, D. (1996). "Institutional Causes of Delay in the Settlement of Legal Disputes". Journal of Law, Economics, and Organization 12, 432–460.

Kessler, D., Meites, T., Miller, G. (1996). "Explaining Deviations from the Fifty-Percent Rule: A Multimodel Approach to the Selection of Cases for Litigation". Journal of Legal Studies 25, 233–259.

Langbein, J.H. (1985). "The German Advantage in Civil Procedure". University of Chicago Law Review 52, 823–866.

Lederman, L. (1999). "Which Cases Go to Trial?: An Empirical Study of Predictors of Failure to Settle". Case Western Reserve Law Review 49, 315–358.

Main, B.G.M., Park, A. (2000). "The British and American Rules: An Experimental Examination of Pre-Trial Bargaining in the Shadow of the Law". Scottish Journal of Political Economy 47, 37–60.

MacCoun, R.J. (1991). "Unintended Consequence of Court Arbitration: A Cautionary Tale From New Jersey". Justice System Journal 14 (2), 229–256.

McKenna, J.A., Wiggins, E.C. (1998). "Empirical Research on Civil Discovery". Boston College Law Review 39, 785–807.

Moller, E. (1996). Trends in Civil Jury Verdicts since 1985. RAND Institute for Civil Justice, Santa Monica, CA.

Moore, K.A. (2001). "Forum Shopping in Patent Cases: Does Geographic Choice Affect Innovation?". North Carolina Law Review 79 (4), 889–938.

Perloff, J.M., Rubinfeld, D.L. (1987). "Settlement in Private Antitrust Litigation". In: Salop, S., White, L. (Eds.), Private Antitrust Litigation. MIT Press, Cambridge, MA, pp. 149–184.

Perloff, J.M., Rubinfeld, D.L., Ruud, P. (1996). "Antitrust Settlements and Trial Outcomes". Review of Economics and Statistics 78, 401–409.

Peterson, M.A., Priest, G.L. (1982). The Civil Jury: Trends in Trials and Verdicts, Cook County Illinois, 1960–79. RAND Institute for Civil Justice, Santa Monica, CA.

P'ng, I.P.L. (1987). "Litigation, Liability, and Incentives for Care". Journal of Public Economics 34 (1), 61–85.

Polinsky, A.M., Rubinfeld, D.L. (1996). "Optimal Awards and Penalties when the Probability of Prevailing Varies among Plaintiffs". RAND Journal of Economics 27, 269–280.

Priest, G.L., Klein, B. (1984). "The Selection of Disputes for Litigation". Journal of Legal Studies 13, 1–55.

Ramseyer, J.M., Rasmusen, E.B. (1997). "Judicial Independence in a Civil Law Regime: The Evidence from Japan". Journal of Law, Economics, and Organization 13, 259–286.

Revesz, R. (2001). "Congressional Influence on Judicial Behavior? An Empirical Examination of Challenges to Agency Action in the D.C. Circuit". New York University Law Review 76, 1100.

Rolph, E., Moller, E., Peterson, L. (1994). Escaping the Courthouse: Private Alternative Dispute Resolution in Los Angeles. RAND Institute for Civil Justice, Santa Monica, CA.

Romano, R. (2002). "The Advantage of Competitive Federalism for Securities Regulation". AEI Press, Washington, DC.

Rosenberg, J.D., Folberg, H.J. (1994). "Alternative Dispute Resolution: An Empirical Analysis". Stanford Law Review 46, 1487–1551.

Rosenfield, A. (1976). "An Empirical Test of Class-Action Settlement". Journal of Legal Studies 5 (1), 113–120.

Rubinfeld, D.L. (1998). "Discovery". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law. Macmillan Reference Limited, London, UK, pp. 608–615.

Shanley, M.G., Peterson, M.A. (1983). Comparative Justice: Civil Jury Verdicts in San Francisco and Cook Counties, 1959–1980. RAND Institute for Civil Justice, Santa Monica, CA.

Shavell, S. (1989). "Sharing of Information Prior to Settlement or Litigation". RAND Journal of Economics 20 (2), 183–195.

Shavell, S. (1996). “Any Frequency of Plaintiff Victory at Trial is Possible”. Journal of Legal Studies 25, 493–501.

Shepherd, G.B. (1999). “An Empirical Study of the Economics of Pretrial Discovery”. International Review of Law and Economics 19 (2), 245–263.

Siegelman, P., Donohue, J.J. III (1995). “The Selection of Employment Disputes for Litigation: Using Business Cycle Effects to Test the Priest-Klein Hypothesis”. Journal of Legal Studies 24 (2), 427–462.

Sloan, F.A., Hoerger, T.J. (1991). “Uncertainty, Information, and the Resolution of Medical Malpractice Claims”. Journal of Risk and Uncertainty 4, 343–363.

Snyder, E.A., Hughes, J.W. (1990). “The English Rule for Allocating Legal Costs: Evidence Confronts Theory”. Journal of Law, Economics & Organizations 6, 345–380.

Solimine, M.E. (1989). “An Economic and Empirical Analysis of Choice of Law”. Georgia Law Review 24, 49–93.

Stookey, J.A. (1990). “Trials and Tribulations, Crises, Litigation, and Legal Change”. Law and Society Review 24 (2), 497–520.

Tabarrok, A., Helland, E. (1999). “Court Politics: The Political Economy of Tort Awards”. Journal of Law & Economics 42 (1), 157–188.

Thiel, S.E. (2000). “Choice of Law and the Home-Court Advantage: Evidence”. American Law & Economics Review 2, 291–317.

Trubek, D.M., Sarat, A., Felstiner, W.L.F., Kritzer, H.M., Grossman, J.B. (1983). “The Costs of Ordinary Litigation”. UCLA Law Review 31, 72–127.

Vidmar, N. (1998). “The Performance of the American Civil Jury: An Empirical Perspective”. Arizona Law Review 40, 849–899.

Viscusi, W.K. (1988). “Product Liability Litigation with Risk Aversion”. Journal of Legal Studies 17, 101–121.

Viscusi, W.K. (2001). “Jurors, Judges, and the Mistreatment of Risk by the Courts”. Journal of Legal Studies 30 (1), 107–142.

Viscusi, W.K., Scharff, R. (1996). “The Structure of the Legal Bargaining Game”. In: Anderson, D.A. (Ed.), Dispute Resolution: Bridging the Settlement Gap. JAI Press, Greenwich, CT.

Waldfogel, J. (1995). “The Selection Hypothesis and the Relationship between Trial and Plaintiff Victory”. Journal of Political Economy 103 (2), 229–260.

Waldfogel, J. (1998). “Reconciling Asymmetric Information and Divergent Expectations Theories of Litigation”. Journal of Law & Economics 41 (2), 451–476.

Willging, T.E., Hooper, L.L., Niemic, R.J. (1996). “An Empirical Analysis of Rule 23 to Address the Rulemaking Challenges”. New York University Law Review 71, 74–185.

Willging, T.E., Shepard, J., Stienstra, D., Miletich, D. (1997). Discovery and Disclosure Practice, Problems, and Proposals for Change: A Case-Based National Survey of Counsel in Closed Federal Civil Cases. Federal Judicial Center, Washington, DC.

Wissler, R.L. (2002). “Court-Connected Mediation in General Civil Cases: What we Know from Empirical Research”. Ohio State Journal on Dispute Resolution 17, 641.

Wittman, D. (1985). “Is the Selection of Cases for Trial Biased?”. Journal of Legal Studies 14, 185–214.

Wittman, D. (1988). “Dispute Resolution, Bargaining, and the Selection of Cases for Trial: A Study of the Generation of Biased and Unbiased Data”. Journal of Legal Studies 17, 313–352.

Chapter 6

THE THEORY OF PUBLIC ENFORCEMENT OF LAW[*]

A. MITCHELL POLINSKY

School of Law, Stanford University, and National Bureau of Economic Research

STEVEN SHAVELL

School of Law, Harvard University, and National Bureau of Economic Research

Contents

[*] Research on this article was supported by the John M. Olin law and economics programs at Stanford Law School and Harvard Law School. An earlier version of portions of this chapter appeared in Polinsky and Shavell (2000a). We are at work on a book-length treatment of public enforcement of law.

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell
© 2007 Elsevier B.V. All rights reserved
DOI: 10.1016/S1574-0730(07)01006-7

Abstract

This chapter surveys the theory of the public enforcement of law—the use of governmental agents (regulators, inspectors, tax auditors, police, prosecutors) to detect and to sanction violators of legal rules. The theoretical core of the analysis addresses the following basic questions: Should the form of the sanction imposed on a liable party be a fine, an imprisonment term, or a combination of the two? Should the rule of liability be strict or fault-based? If violators are caught only with a probability, how should the level of the sanction be adjusted? How much of society's resources should be devoted to apprehending violators? A variety of extensions of the central theory are then examined, including: activity level; errors; the costs of imposing fines; general enforcement; marginal deterrence; the principal-agent relationship; settlements; self-reporting; repeat offenders; imperfect knowledge about the probability and magnitude of sanctions; corruption; incapacitation; costly observation of wealth; social norms; and the fairness of sanctions.

Keywords

public enforcement of law, fines, imprisonment, strict liability, fault-based liability, probability of detection, errors, general enforcement, marginal deterrence, settlements, self-reporting, repeat offenders, fairness of sanctions, norms

JEL classification: D23, D62, D63, H23, H26, K14, K42, L51

1. Introduction

Public enforcement of law—the use of governmental agents to detect and to sanction violators of legal rules—is a subject of obvious importance. Police and prosecutors endeavor to solve crimes and to punish criminals, regulators attempt to control violations of environmental, safety, consumer protection, and financial disclosure laws, and agents of the Internal Revenue Service seek to enforce the tax code.

The earliest economically-oriented writing on the subject of public law enforcement dates from the eighteenth century contributions of Montesquieu (1748), Beccaria (1767), and, especially, Bentham (1789). Curiously, after Bentham, the subject of law enforcement lay essentially dormant in economic scholarship until the late 1960s, when Becker (1968) published a highly influential article. Since then, several hundred articles have been written on the economics of law enforcement.[1]

The main purpose of our chapter is to present the economic theory of public law enforcement in a systematic and comprehensive way.[2] The theoretical core of our analysis addresses the following basic questions: Should the form of the sanction imposed on a liable party be a fine, an imprisonment term, or a combination of the two? Should the rule of liability be strict or fault-based? If violators are caught only with a probability, how should the level of the sanction be adjusted? How much of society's resources should be devoted to apprehending violators?

The chapter is outlined as follows. We begin in section 2 by considering the rationale for public enforcement of law, that is, by asking why society cannot rely exclusively on private enforcement of law to control undesirable behavior. We then state the problem of public enforcement of law in general terms in section 3. In sections 4 through 6, we analyze strict liability and fault-based liability when enforcement is certain, first considering monetary sanctions, next non-monetary sanctions, and then the two together. In sections 7 through 9, we perform the same analysis when enforcement is uncertain because it is costly. In section 10 we summarize the theory, and in section 11 we discuss enforcement practices in the light of the theory we have reviewed.

We subsequently examine a variety of extensions of our core analysis in sections 12 through 28. These concern mistake, the costs of imposing sanctions, marginal deterrence, the settlement process, self-reporting of violations, corruption of law enforcers, and the fairness of sanctions, among other topics. We conclude in section 29.

[1] See, for example, the references cited in Bouckaert and De Geest (1992, pp. 504–526), Garoupa (1997), Mookherjee (1997), and Polinsky and Shavell (2000a).

[2] For other surveys of the theory of public enforcement, see Garoupa (1997) and Mookherjee (1997). For surveys of empirical research on law enforcement and crime, see Eide (2000) and Levitt and Miles (2007), and for a survey of empirical research on enforcement of environmental regulation, see Cohen (1999, pp. 78–95).

2. Why public enforcement rather than private enforcement?

Before proceeding, we should comment on the rationale for public, as opposed to private, law enforcement, where, by the latter, we mean the bringing of suits by victims of harm or those threatened by harm. An important element of the justification for private enforcement concerns information about the identity of violators. When victims of harm naturally possess knowledge of who injured them, allowing private suits for harm will motivate victims to initiate legal action and thus will harness the information they have for purposes of law enforcement. This may help to explain why, for example, the enforcement of contract law and tort law is primarily private in nature: a victim of a contract breach obviously knows who committed the breach, and a victim of a tort usually knows who the tortfeasor was. When, however, victims cannot easily identify who injured them, it may be desirable for public enforcement to be employed.

For public enforcement to be preferred in such circumstances, one still needs to explain why society cannot rely on rewards of some type to private parties other than victims (such as friends of violators or entrepreneurial private enforcers) to supply information and otherwise help in detecting violators. A difficulty with reliance on private enforcement of this sort is that if a reward is available to everyone, there might be wasteful effort devoted to finding violators (akin to excessive effort to catch fish from a common pool). Another problem is that private parties may find it hard to capture fully the benefits of developing expensive, but socially worthwhile, information systems to aid enforcement (such as computerized databases of fingerprint records). An additional obstacle to private enforcement is that force may be needed to gather information, capture violators, and prevent reprisal, yet the state frequently, if not usually, will not want to permit private parties to use force. For the preceding reasons, public enforcement often will be favored when effort is required to identify and apprehend violators.[3]

3. The general problem of public law enforcement

The general problem of public law enforcement may be viewed as one of maximizing social welfare. By social welfare, we refer to the benefits that individuals obtain from their behavior, less the costs that they incur to avoid causing harm, the harm that they do cause, the cost of catching violators, and the costs of imposing sanctions on them (including any costs associated with risk aversion). We will be explicit below about the definition of social welfare in the various contexts that we consider.

[3] The differences between public and private enforcement have been discussed by Becker and Stigler (1974), Landes and Posner (1975), and Polinsky (1980a); see also Shavell (1993) and Friedman (1995). In this chapter, we assume for simplicity that public enforcement is the exclusive means of enforcement, even though in practice private parties sometimes play a complementary role by supplying information to enforcement authorities and by bringing private suits. We also abstract from private parties' efforts to protect themselves from harm (and how such efforts might relate to public enforcement), though we mention this issue in the conclusion.

The state has four major policy choices to make in undertaking law enforcement. One is about the sanctioning rule. The rule could be *strict* in the sense that a party is sanctioned whenever he has been found to have caused harm (or expected harm). Alternatively, the rule could be *fault-based*, meaning that a party who has been found to have caused harm is sanctioned only if he failed to obey some standard of behavior or regulatory requirement.

A second choice of the state concerns the form of the sanction: monetary versus non-monetary (both may be employed together). We focus on imprisonment as the primary type of non-monetary sanction and we assume that monetary sanctions are socially less costly to employ than imprisonment.

A third choice involves the magnitude of the sanction.

And the fourth choice concerns the probability of detecting offenders and imposing sanctions. This probability depends on the public resources devoted to finding violators and proving that they are liable.[4]

Part A: Basic theory when enforcement is certain

4. Monetary sanctions

In this section we analyze the optimal magnitude of monetary sanctions—which we call fines—assuming that enforcement is certain. We consider the two basic forms of liability, beginning with strict liability.

Suppose that individuals would obtain a gain from committing a harmful act, where the gain varies among them. If an individual does commit the act, he will have to pay a fine because he is strictly liable. Let

$g =$ gain an individual obtains if he commits the harmful act;

$z(g) =$ density of gains among individuals;

$h =$ harm caused by an individual if he commits the harmful act;[5]

$f =$ fine; and

$w =$ level of wealth of an individual.

The gain could be a literal benefit, for instance, the utility obtained from taking food from a cabin in the wilderness, or the savings from not investing in some precaution, such as not obtaining pollution control equipment.[6] The fine cannot exceed an individual's level of wealth, which is assumed to be the same for everyone.

[4] The framework for studying public law enforcement employed in this chapter derives, in many respects, from Bentham (1789). Becker (1968) first stated the enforcement problem in formal economic terms and added the choice of the probability of detection to Bentham's expression of the enforcement problem.

[5] If the harm is uncertain, h can be interpreted as expected harm instead of actual harm; see section 12.

[6] For simplicity, we assume that the gain is not itself available to pay the fine. This assumption would hold if the gain were non-monetary, as in the example of the utility benefit from taking food from a cabin. In many circumstances, the assumption would not be fitting, but the complications introduced by considering how the gain enhances an individual's ability to pay a fine would be distracting.

An individual will commit the harmful act if and only if his gain from doing so exceeds the fine:[7] $g > f$.

Social welfare equals the gains individuals obtain from committing the harmful act less the harm caused.[8] Social welfare is not directly affected by the imposition of fines because the payment of a fine is assumed to be a socially costless transfer of money.[9] Since the individuals who commit the harmful act are those whose gains exceed the fine, social welfare is

$$\int_{f}^{\infty} (g - h)z(g)dg. \tag{1}$$

The enforcement authority's problem is to maximize social welfare by choosing the fine f. We use an asterisk to denote the optimal value of the fine (and other variables below). It is clear from (1) that

$$f^* = h, \tag{2}$$

assuming that such a fine is feasible. Hence, individuals commit the harmful act if and only if their gain exceeds the harm, which is first-best behavior. Note, however, that there will be underdeterrence if the individual's wealth is less than h, in which case the optimal fine equals the individual's wealth.

Now consider the fault-based sanctioning rule, in which an individual who causes harm is sanctioned only if he failed to obey some standard of behavior. In the present framework, we assume that if an individual commits a harmful act, his gain must equal or exceed some threshold level of gain in order for him to be free of liability; otherwise, he is considered to be at fault. Let

$\hat{g}$ = threshold level of gain under the fault-based sanctioning rule.

If an individual's gain is less than $\hat{g}$ he will commit the harmful act whenever $g > f$, while if his gain equals or exceeds $\hat{g}$, he will commit the harmful act regardless of his gain. In other words, the individual will commit the harmful act under fault-based liability if $g > \min(f, \hat{g})$. Thus, social welfare is

$$\int_{\min(f,\hat{g})}^{\infty} (g - h)z(g)dg. \tag{3}$$

[7] We assume for simplicity that he does not commit the harmful act if he is indifferent.

[8] Some writers have questioned whether gains from committing harmful acts should necessarily be credited in social welfare; see, for example, Stigler (1970, p. 527) and Lewin and Trumbull (1990). If the gains from some type of harmful conduct were excluded from social welfare, the main consequence for our analysis would be that, for this type of conduct, society would want to achieve greater, possibly complete, deterrence. That, in turn, would tend to make a higher sanction and a higher probability of detection desirable.

[9] In practice, of course, some costs are incurred in collecting fines, such as the cost of identifying and confiscating the individual's assets if the individual resists paying the fine. We discuss the implications of such costs in section 16.

The enforcement authority's problem is to maximize (3) by choosing the standard $\hat{g}$ and the fine f. It is straightforward to see that

$$f^* = \hat{g}^* = h \tag{4}$$

results in first-best behavior. In particular, if $g < h$, the individual would be at fault if he committed the harmful act, and would bear a fine equal to h, so he would not commit it. If $g \geq h$, he would not be at fault if he committed the harmful act and therefore would not have to pay anything, so he would commit the act. Note that f^* and $\hat{g}^*$ are not unique: first-best behavior also can be achieved by setting $\hat{g} > h$, with $f = h$, or by keeping $\hat{g} = h$ and setting $f > h$.

Although we have seen that the first-best outcome can be achieved under both strict liability and fault-based liability (when a fine equal to h is feasible), the two forms of liability differ in the information required to implement them. To apply strict liability, the state only needs to know the harm. Under fault-based liability, however, the state needs to know more: it also needs to ascertain the gain of the individual (to determine whether he was at fault).

To illustrate strict and fault-based liability, suppose that a firm contemplates whether to discharge a pollutant that would cause harm h, rather than to transport it to a waste disposal site at cost g. Under strictly-imposed fines, the firm would incur a fine of h if it pollutes, so would not do so unless its savings from not transporting the waste, g, is greater than h. Under fault-based fines, a polluting firm would incur a fine of h only if the cost of transporting the waste g is less than h; thus, the firm would pollute only if $g \geq h$.

5. Nonmonetary sanctions

In this section we analyze the optimal magnitude of nonmonetary sanctions—which we assume to be prison sentences[10]—when enforcement is certain. Let

$$s = \text{length of prison sentence, where } s \text{ is in } [0, s_M]; \text{ and}$$
$$d(s) = \text{disutility from prison sentence of length } s;\ d(0) = 0;\ d'(s) > 0.$$

The maximum sentence s_M can be interpreted as life imprisonment, and is assumed to be the same for everyone.

Under strict liability, if an individual commits the harmful act, he will bear a prison sentence.[11] Thus, he will commit the act if and only if his gain from doing so exceeds the disutility of the prison sentence: $g > d(s)$.

[10] It will be clear that most of what we have to say about imprisonment here and in subsequent sections would carry over, with only slight modification, to other forms of non-monetary sanctions, such as probation, electronic monitoring, community service, or, in the extreme, the death penalty.

[11] Strict liability is in fact an unusual form of liability when the sanction is imprisonment. (An example of a strict liability criminal offense is the serving of liquor by a bar to underage individuals.) As will be seen, fault-based liability tends to be superior to strict liability when the sanction is imprisonment.

Social welfare equals the gains individuals obtain from committing the harmful act, less the harm caused, and less the cost of imposing prison sentences. The cost of prison sentences is the sum of the disutility suffered by the sanctioned individuals and the cost to the state of maintaining a prison system. Let

$c =$ cost to the state per unit of prison sentence; $c > 0$.

Thus, social welfare is

$$\int_{d(s)}^{\infty} [g - h - (d(s) + cs)]z(g)dg. \tag{5}$$

The enforcement authority's problem is to maximize social welfare by choosing the prison sentence s. From (5), the first-order condition can be written as

$$(h + cs)(dZ(d(s))/ds) = \int_{d(s)}^{\infty} [d'(s) + c]z(g)dg, \tag{6}$$

where $Z(\cdot)$ is the cumulative distribution function for $z(\cdot)$. The left-hand side is the marginal benefit of raising s: $h + cs$ is the social gain from deterring the marginal individuals, while $dZ(d(s))/ds$ is the number of such individuals deterred.[12] The right-hand side is the marginal cost of raising s: individuals who are not deterred incur additional disutility $d'(s)$ and cause the state to incur additional imprisonment costs c.

The optimal prison sentence could lead to either underdeterrence or overdeterrence due to the costs of prison sentences. To see why, suppose the sentence were such that individuals committed the harmful act if and only if their gain equaled the harm. If the sentence were raised above this level, some individuals would be deterred, which would reduce prison costs, but those who would not be deterred would bear longer sentences. Depending on which effect is stronger, it may be desirable to raise, or lower, the sentence, leading to overdeterrence or underdeterrence. Note that a marginal change of the sentence from the initial level does not affect the gains net of the harm because the marginal individuals are those whose gains equal the harm.[13]

Regardless of whether the optimal prison sentence causes underdeterrence or overdeterrence, the strict sanctioning rule does not achieve the first-best outcome because it leads to the imposition of costly sanctions.

Now consider the fault-based sanctioning rule, with a standard $\hat{g}$. Analogously to the case of fines, an individual will commit the harmful act under fault-based liability if $g > \min(s, \hat{g})$. If $s < \hat{g}$, then there will be some individuals who commit the harmful

[12] The disutility of imprisonment and the benefit from committing the harmful act do not appear in (6) because they offset each other for the marginal person.

[13] The point that underdeterrence or overdeterrence may result when the prison sentence is optimally chosen was made by Polinsky and Shavell (1984, p. 94). See also Kaplow (1990b).

act who will be found at fault, those for whom $s < g < \hat{g}$. Then social welfare would be

$$\int_{s}^{\infty}(g-h)z(g)dg - \int_{s}^{\hat{g}}(d(s)+cs)z(g)dg. \tag{7}$$

If $s \geq \hat{g}$, then social welfare is

$$\int_{\hat{g}}^{\infty}(g-h)z(g)dg. \tag{8}$$

The enforcement authority's problem is to maximize (7) or (8) by choosing the standard $\hat{g}$ and the sentence s. It is straightforward to see that

$$s^* = \hat{g}^* = h \tag{9}$$

results in the first-best outcome. First observe that first-best behavior will occur. In particular, if $g < h$, the individual would be at fault if he committed the harmful act, and bear a sentence equal to h, so he would not commit it. If $g \geq h$, he would not be at fault if he committed the harmful act and therefore would not bear any sentence, so he would commit the act. Second, given (9), costly sanctions will not be imposed because individuals will never choose to be at fault. Hence, the first-best outcome is achieved. Note that s^* and $\hat{g}^*$ are not unique: the first-best outcome also can be achieved by keeping $\hat{g} = h$ and setting $s > h$.

The preceding discussion shows that when sanctions are costly, the fault-based sanctioning rule is superior to the strict sanctioning rule. Not only does the fault-based rule lead to first-best deterrence, it does so without anyone actually incurring a costly sanction.[14]

6. Combined sanctions

In this section we consider the optimal mix of fines and imprisonment sanctions when they can be used together. Under the strict sanctioning rule, social welfare in this case is

$$\int_{f+d(s)}^{\infty}[g-h-(d(s)+cs)]z(g)dg. \tag{10}$$

First observe that it cannot be optimal to employ a prison sentence unless the fine has been set as high as possible, equal to individuals' wealth level w. To see why, suppose otherwise, that $f < w$ and $s > 0$. Then it would be possible to raise f and lower s so as to keep $f + d(s)$ constant, thereby leaving behavior unaffected but lowering the cost of imprisonment (see (10)).

[14] This advantage of the fault-based rule was originally emphasized in Shavell (1987b), which also considered the possibility of error, and thus of the bearing of sanctions. On the latter issue, see section 15 below.

Whether it is optimal to use a prison sentence in addition to a fine depends on whether a fine alone is sufficient to achieve first-best deterrence. If wealth levels are high enough, if $w \geq h$, then a fine equal to harm is feasible and there would be no need for a prison sentence. If, however, $w < h$, then relying on fines alone would lead to underdeterrence and it might be desirable to employ a prison sentence in addition to the maximal fine of w. To see whether a prison sentence would be desirable in these circumstances, consider the derivative of (10) with respect to the sentence s when $f = w$:

$$[h - w + cs][dZ(w + d(s))/ds] - \int_{w+d(s)}^{\infty} [d'(s) + c]z(g)dg. \tag{11}$$

It follows that the condition for a positive prison sentence to be optimal is that

$$[h - w][dZ(w)/ds] > \int_{w}^{\infty} [d'(0) + c]z(g)dg. \tag{12}$$

The left-hand side of (12) is the value of deterring the marginal individuals, while the right-hand side is the marginal cost of imposing prison sentences on individuals who are not deterred. If a positive prison sentence is socially desirable, it is determined from the first-order condition associated with (11).

Next consider the optimal mix of fines and imprisonment under the fault-based sanctioning rule. The key point is that, unlike under strict liability, it is always desirable to employ prison sentences to obtain compliance with the fault standard if fines alone are inadequate to do so. This is because, as noted above, sanctions are not actually imposed when the fault standard is complied with, so there is no social cost from using the threat of prison sentences to obtain compliance. Thus, any combination of fines and prison sentences that induces potential offenders to comply with the fault standard is optimal.

Part B: Basic theory when enforcement is uncertain

In this part we investigate the level of enforcement resources that the state should devote to detecting offenders. We assume that the higher the level of expenditures on enforcement, the greater is the probability of detection. Let

e = enforcement expenditures by the state; and

$p(e)$ = probability of detection given e; $p'(e) > 0$; $p''(e) < 0$.

We derive the optimal probability of detection[15] and, along with it, the optimal magnitude of sanctions.[16]

[15] We implicitly assume that enforcement expenditures e determine a single probability of detection. Thus, we do not consider an issue identified by Lando and Shavell (2004), that it may be advantageous to concentrate enforcement resources on a subset of potential offenders (for example, by auditing taxpayers whose last names begin with certain letters) rather than to spread enforcement resources evenly.

[16] Although we assume that the probability of detection can be set independently of the level of sanctions, the two might be connected. This is because high sanctions may lead juries to be less likely to convict defendants,

7. Monetary sanctions

In this section we analyze the optimal probability and magnitude of fines, first assuming that individuals are risk neutral and then that they are risk averse.

7.1. The risk-neutral case

Under the strict liability rule, an individual will commit the harmful act if and only if his gain from doing so exceeds the expected fine: $g > pf$. Social welfare, which now reflects the enforcement expenditures of the state, is

$$\int_{p(e)f}^{\infty} (g - h)z(g)dg - e. \tag{13}$$

The enforcement authority's problem is to maximize (13) by choosing enforcement expenditures e (and thus the probability of detection p), as well as the level of the fine f.

Before considering the complete problem, suppose that enforcement expenditures are fixed, resulting in the probability of detection p. It is obvious that if $pf = h$, namely, the expected fine equals the harm, (13) will be maximized over f since the first-best outcome will be achieved. In other words,

$$f^* = h/p, \tag{14}$$

assuming that such a fine is feasible. Individuals then commit the harmful act if and only if their gain exceeds the harm, which is first-best behavior.[17] Note, however, that there will be underdeterrence if individuals' wealth is less than h/p, in which case the optimal fine equals their wealth.

Now suppose that both enforcement expenditures and the fine are chosen by the state.[18] Then the optimal fine is maximal: $f^* = w$. To demonstrate this, suppose that f is less than w. Then f can be raised and e lowered so as to keep $p(e)f$—the level of deterrence—constant. Because the behavior of individuals is unaffected but enforcement expenditures fall, social welfare rises (the first term in (13) does not change but e is lower). Hence, the optimal f cannot be less than w. In other words, because any particular level of deterrence can be achieved with different combinations of the fine and the probability of detection, society should employ the highest possible fine and

or may induce individuals to engage in greater efforts to avoid detection; on these points, see Andreoni (1991) and Malik (1990), respectively. See also Bar-Gill and Harel (2001) for a discussion of how the level of crime affects both the probability and magnitude of sanctions.

[17] The general formula (14), or its equivalent, was put forward by Bentham (1789, p. 173), was emphasized by Becker (1968), and has been noted by many others since then.

[18] Consideration of this issue originated with Becker (1968); early writers on law enforcement did not examine the issue of the choice of enforcement effort.

a correspondingly low probability of detection in order to economize on enforcement expenditures.[19]

We next show that the optimal probability of detection is such that the expected fine is less than the harm, $p(e^*)w < h$—that is, some degree of underdeterrence is desirable. Observe that the first-order condition determining optimal enforcement expenditures e^* is

$$[h - p(e)w][dZ(p(e)w)/de] = 1. \tag{15}$$

The left-hand side is the marginal social benefit of the deterrent effect from a higher probability of detection. The right-hand side is the marginal cost of greater spending on enforcement. It follows from (15) that $p(e^*)w < h$. To explain this result, suppose that p were such that $pw = h$. Then there would be no first-order loss of social welfare from lowering p because the individuals who would be induced to engage in the harmful activity would obtain gains equal to harm. But enforcement costs would be saved, making it desirable to lower the probability. How much p should be lowered depends on the resulting savings in enforcement expenses compared to the net social costs of underdeterrence.[20]

Under fault-based liability, analogues of the above conclusions hold. The optimal fine is maximal and the optimal probability of detection is such that the expected fine is less than the harm. Moreover, the optimal fault standard is less than the first-best standard. The explanation for these results is essentially that given above. The fine is maximal in order to reduce the probability of detection and thereby save enforcement costs, and some underdeterrence is desirable because this also allows savings in enforcement costs.

7.2. *The risk-averse case*

Now suppose that individuals are risk averse and that social welfare is the sum of expected utilities of individuals (in the risk-neutral case, social welfare was equivalent to the sum of utilities). For convenience, assume that the risk of being harmed is the same for everyone and that individuals buy insurance against harm, paying a premium equal to the expected harm. An individual's wealth is his initial wealth, less the taxes he pays,

[19] Although the general point that a low probability-high fine combination conserves enforcement costs is due to Becker (1968), he did not formally consider bounds on fines (and much of his analysis implicitly presumes that the optimal fine is not maximal). Carr-Hill and Stern (1979, pp. 300–304) and Polinsky and Shavell (1979, pp. 883–884) observed that Becker's argument implies that the optimal fine is equal to its upper bound. Many scholars have noted the unrealism of this result and have introduced additional considerations that imply that less-than-maximal fines are optimal. We will discuss several important factors of this type, including risk aversion, general enforcement, and marginal deterrence. See also Andreoni (1991), Bebchuk and Kaplow (1992, 1993), Malik (1990), and Polinsky and Shavell (1991, 2000b) for discussion of other such considerations.

[20] The point of this paragraph—that some underdeterrence is optimal—was first made by Polinsky and Shavell (1984).

less the expected harm he suffers, and less the fine if he has to pay it. Let

$$
\begin{aligned}
U(\cdot) &= \text{ utility of wealth; } U \text{ is concave in wealth;}\\
t &= \text{tax; and}\\
\lambda &= \text{fraction of population that commits the harmful act.}
\end{aligned}
$$

Both t and λ are endogenous.

An individual will commit the harmful act if $g + (1 - p)U(w - t - \lambda h) + pU(w - t - \lambda h - f)$ is greater than $U(w - t - \lambda h)$, or equivalently, if

$$g > p[U(w - t - \lambda h) - U(w - t - \lambda h - f)]. \qquad (16)$$

Note that we are implicitly assuming that the gain g is non-monetary.[21] The condition (16) implicitly determines the fraction of the population λ that commits the harmful act. The tax t is such that the government breaks even; hence t equals the enforcement cost e less the fine revenue collected λpf.

Social welfare, the sum of individuals' expected utilities, equals

$$
\begin{aligned}
&(1 - \lambda)U(w - t - \lambda h)\\
&\quad + \int_{p[U(w-t-\lambda h)-U(w-t-\lambda h-f)]}^{\infty} [g + (1 - p)U(w - t - \lambda h)\\
&\quad + pU(w - t - \lambda h - f)]z(g)dg - e, \qquad (17)
\end{aligned}
$$

since the individuals who commit the harmful act are those whose gains exceed the expected disutility of the fine.

Suppose initially that the probability of detection is fixed. Then the optimal fine in the risk-averse case tends to be lower than that in the risk-neutral case for two reasons. First, lowering the fine reduces the bearing of risk by individuals who commit the harmful act. Second, because risk-averse individuals are more easily deterred than risk-neutral individuals, the fine does not need to be as high to achieve any desired degree of deterrence.[22]

Now consider choosing both the probability and magnitude of fines. The optimal fine generally is not at its maximum when individuals are risk averse. This is because the use of a very high fine would impose a substantial risk-bearing cost on individuals who commit the harmful act. More precisely, reconsider the argument employed in the risk-neutral case. If f is less than the maximal fine (now $w - t - \lambda h$), it still is true that f can be raised and e lowered so as to keep deterrence constant. But due

[21] If the gain were monetary, then the condition would become $(1 - p)U(w + g - t - \lambda h) + pU(w + g - t - \lambda h - f) > U(w - t - \lambda h)$. The qualitative nature of the results in this section would not be affected.

[22] It is possible, however, that the optimal fine is higher in the risk-averse case than in the risk-neutral case, for the following reason. A way to reduce the bearing of risk is to deter more individuals from committing the harmful act, for then fewer individuals will be subject to the risk of the fine. See Kaplow (1992).

to risk aversion, the probability of detection that maintains deterrence falls *more* than proportionally, implying that the expected fine, and therefore fine revenue, falls. This reduction in fine revenue reflects the disutility caused by imposing greater risk on risk-averse individuals. If individuals are sufficiently risk averse, the decline in fine revenue associated with greater risk bearing could more than offset the savings in enforcement expenditures from reducing the probability of detection, implying that social welfare would be lower.[23]

In effect, when individuals are risk averse, fines become a socially costly sanction (reflected in an increase in taxes) rather than a mere transfer of wealth. The more risk averse individuals are, the better it is to control their behavior by using a lower fine and a higher probability of detection, even though this raises enforcement costs.

As in the risk-neutral case, there is a reason when individuals are risk averse to reduce enforcement costs by setting the probability such that some individuals will commit the harmful act even though their gain is less than the harm—meaning that there will be some underdeterrence.[24]

Under fault-based liability, the conclusions are different from those under strict liability. The differences are due to the fact that individuals are induced to comply with the fault standard and therefore do not bear risk. Consequently, the results are the same as those in the risk-neutral case: the optimal fine is maximal; the optimal probability is relatively low; and the optimal fault standard is such that there is some underdeterrence.

Because, as just emphasized, fines are not actually imposed under fault-based liability, fault-based liability is superior to strict liability, under which risk is borne. This advantage of fault-based liability is analogous to the advantage of fault-based liability over strict liability when imprisonment is used—a costly sanction is not actually imposed.

Another advantage of fault-based liability is that it may result in lower enforcement expenditures than under strict liability. Specifically, because fines are not imposed under fault-based liability, it becomes desirable to use a high (maximal) fine, which allows a relatively low probability of detection to be employed.

Of course, in reality, as we will be discussing below, mistakes may occur under fault-based liability, resulting in the imposition of risk. To the extent that risk exists under fault-based liability, the main result obtained under strict liability—that fines should not be maximal—carries over to fault-based liability.

[23] The point that the optimal fine may be less than maximal when individuals are risk averse was proved initially by Polinsky and Shavell (1979) in a model with two levels of gain. See also Kaplow (1992), who demonstrates in an example that the fine may be less than maximal. It can be shown in the general model under discussion here that the optimal fine must be less than maximal if the cost of raising the probability of detection is sufficiently small (given that the wealth of individuals is not too low). The idea of the proof is that, if the cost of raising the probability were zero, the optimal probability would be one and the optimal fine less than maximal (equal to the harm). The conclusion follows by a continuity argument.

[24] It also is possible, however, that overdeterrence would be optimal. The reason is that the imposition of risk can be reduced by discouraging individuals from engaging in the harmful activity.

8. Nonmonetary sanctions

In this section we analyze the optimal probability and magnitude of prison sentences, first assuming that individuals are risk neutral, then that they are risk averse, and finally that they are risk preferring. The last case is of particular relevance for imprisonment sanctions, as will be explained.

8.1. The risk-neutral case

We assume here that $d(s) = s$, that is, that the disutility of imprisonment rises in proportion to the length of the sentence. This implies that individuals are indifferent between a sure sentence of s and an uncertain sentence with a mean of s. Thus, individuals display a risk-neutral attitude towards imprisonment sentences.

Under the strict liability rule, an individual will commit the harmful act if and only if his gain exceeds the expected sentence: $g > ps$. Social welfare is

$$\int_{p(e)s}^{\infty} (g - h - p(e)(s + cs))z(g)dg - e. \tag{18}$$

The enforcement authority's problem is to maximize (18) by choosing enforcement expenditures e and the length of the sentence s.

In this case the optimal sentence is maximal: $s^* = s_M$. As seen above, when the sanction is a fine, the optimal sanction is maximal. This is also true here even though the sanction is costly to impose. To demonstrate that $s^* = s_M$, suppose that s is less than s_M. Then s can be raised and e lowered so as to keep $p(e)s$ constant. The behavior of individuals is unaffected because $p(e)s$ has not changed. The social cost of imprisonment also is unaffected because $p(e)(s + cs)$ is constant. In other words, although the sentence is higher, proportionally fewer individuals are imprisoned. But because enforcement expenditures fall, social welfare rises. Hence, the optimal s equals s_M.

The optimal probability of detection is such that the expected prison sentence could lead to either underdeterrence or overdeterrence. This is essentially for the reasons discussed above (see section 5) when the probability of detection was fixed at one. Here, however, there is an additional factor favoring underdeterrence, namely that by lowering the probability, enforcement resources are saved.

Under fault-based liability, first observe that it must be optimal to have compliance with the fault standard. If the expected sentence were less than the standard, so that some individuals would choose to violate the standard and bear the expected sentence, then it would be optimal to lower the standard to the expected sentence. For then, the behavior of individuals would be the same, but the cost of imprisonment would be avoided.

Next observe that the expected sentence must equal the standard, rather than be higher. Otherwise, the probability of detection could be lowered without affecting behavior, but enforcement costs would be saved.

It is clear, too, that the optimal sentence must be maximal. This is for the now familiar reason that, otherwise, the sentence could be raised and the probability of detection lowered proportionally, without affecting behavior, but saving enforcement expenditures (imprisonment costs are zero because of compliance with the standard).[25] Given $s^* = s_M$, the optimal probability satisfies $p(e)s_M = \hat{g}$ because of our observation in the previous paragraph.

The optimal standard $\hat{g}^*$ is less than the harm h; in other words, there is some underdeterrence. This is true for a reason already discussed: by lowering the standard, enforcement costs are saved, but there is no first-order effect on social welfare due to more individuals committing the harmful act because their gain equals the harm.

Finally, as we previously emphasized, fault-based liability possesses the advantage over strict liability that costly sanctions are not actually imposed (in the absence of mistakes). Moreover, because the optimal sentence is maximal for this reason, a low probability of detection can be used.

8.2. *The risk-averse case*

We assume now that the disutility of the sentence $d(s)$ rises more than in proportion to the sentence. This could occur because of an increasing desire of a prisoner to join the outside world or a growing distaste for the prison environment as time in jail increases. This assumption implies that individuals prefer a sure sentence of s to an uncertain sentence with a mean of s. In other words, individuals are risk-averse in imprisonment sentences.

Under the strict liability rule, an individual will commit the harmful act if and only if his gain exceeds the expected disutility of the sentence, $g > pd(s)$, and social welfare is

$$\int_{p(e)d(s)}^{\infty} g - h - p(e)(d(s) + cs))z(g)dg - e. \tag{19}$$

The optimal sentence is again maximal: $s^* = s_M$. The reasons given in the risk-neutral case are reinforced here. As s is raised and e lowered so as to keep $p(e)d(s)$ and the behavior of individuals constant, p declines proportionally more than s rises because of individuals' risk aversion. In other words, ps declines, which implies that the public cost of imprisonment pcs falls.[26] Hence, social welfare rises both because the cost of enforcement declines and the cost of imprisonment declines.

The optimal probability of detection is such that, as in the risk-neutral case, the expected prison sentence could lead to underdeterrence or overdeterrence.

[25] Note, too, that the expected sentence remains constant, so even if imprisonment costs were borne, they would not change.

[26] Note the contrast with the case of risk aversion in fines, in which the decline in the expected sanction pf meant a decline in revenue to the state and thus an increase in taxes. Here the decline in the expected sanction ps means a decline in expenses to the state and thus a decrease in taxes.

Under fault-based liability, the results are essentially the same as in the risk-neutral case because, if the fault standard is complied with, risk is not borne. Thus, it is optimal to have compliance with the fault standard; the expected disutility of the sentence must equal the standard; the optimal sentence is maximal; there is some underdeterrence; and fault-based liability is superior to strict liability. A difference, however, is that the probability of detection needed to obtain compliance with the fault standard can be lowered because individuals are risk averse.

8.3. The risk-preferring case

Finally, suppose that the disutility of the sentence $d(s)$ rises less than in proportion to the sentence. This could occur because the disutility from the stigma of being in jail might be substantial from having spent even a short amount of time there, but not increase much with the length of imprisonment.[27] Individuals' discounting of the future disutility of imprisonment also makes earlier years of imprisonment more important than later ones. The present assumption implies that individuals prefer an uncertain sentence with a mean of s to a sure sentence of s; individuals are risk-preferring in imprisonment sentences.

Under the strict liability rule, social welfare is again given by (19). The optimal sentence, however, might be less than maximal: $s^* \leq s_M$. Now, when the sentence is raised, the probability that maintains deterrence cannot be lowered proportionally, implying that the expected prison term rises. Because the resulting increased cost to the public of imposing imprisonment sanctions might exceed the savings in enforcement expenditures from lowering the probability, the optimal prison term might not be maximal.[28]

Under fault-based liability, the results are again essentially the same as in the risk-neutral case.

9. Combined sanctions

Under the strict sanctioning rule, as we explained above, it never is optimal to employ a prison sentence unless the fine has been set as high as possible, since fines are socially cheaper sanctions. Whether it is optimal to use a prison sentence in addition to the maximal fine depends on the extent of underdeterrence that would result if fines were used alone, and the social cost of imprisonment.

Under the fault-based sanctioning rule, the key point is that it is always desirable to employ the maximal prison sentence in addition to the maximal fine, since neither sanction is actually imposed. By using maximal sanctions, the probability of detection can be set at a low level, thereby saving enforcement costs.

[27] Also, the first years of imprisonment may create special disutility due to brutalization of the prisoner.

[28] The results in this section were first presented by Polinsky and Shavell (1999) (although Shavell (1991b) notes the result in the case of risk neutrality).

Part C: Basic theory summarized and compared to practice

10. Summary of the basic theory

In this section we summarize the main points from parts A and B:

(a) When the probability of detection of a harmful act is taken as fixed and individuals are risk neutral, the optimal fine is the harm divided by the probability of detection, for this results in an expected fine equal to the harm. However, the risk aversion of individuals tends to lower the level of the optimal fine.

(b) When the probability of detection can be varied, relatively high sanctions may be optimal, for this allows a relatively low probability to be employed and thereby saves enforcement costs. Indeed, the optimal fine is maximal if individuals are risk neutral in wealth, and the optimal imprisonment term is maximal if individuals are risk neutral or risk averse in imprisonment. Optimal sanctions might not be maximal, however, when individuals are risk averse in wealth or risk preferring in imprisonment, both plausible assumptions, although the motive to set sanctions at relatively high levels in order to reduce enforcement costs still applies.[29]

(c) Optimal enforcement tends to be characterized by some degree of underdeterrence relative to first-best behavior, because allowing underdeterrence conserves enforcement resources. More precisely, by lowering the probability of detection slightly from a level that would lead to first-best behavior, the state reduces enforcement costs, and although more individuals commit the harmful act, these individuals do not cause social welfare to decline substantially because their gains are approximately equal to the harm.

(d) The use of fines should be exhausted before resort is made to the costlier sanction of imprisonment.

(e) An advantage of fault-based liability over strict liability is that sanctions that are costly to impose—imprisonment, and fines when individuals are risk averse—are imposed less often under the former rule. Under fault-based liability, individuals generally are induced (in the absence of mistakes) to obey fault standards, and therefore ordinarily do not bear sanctions. Under strict liability, however, individuals who cause harm are sanctioned whenever they are caught.

(f) An advantage of strict liability over fault-based liability is that the former is easier to apply. Strict liability requires the state to determine only the harm done, whereas fault-based liability requires the state to ascertain optimal behavior (in order to set the fault standard) and to observe actual behavior (in order to apply the standard).

11. Theory versus practice

Having reviewed the basic theory of public enforcement of law, we briefly comment on the relationship between optimal enforcement and enforcement in practice.

[29] There are other reasons why optimal sanctions might not be maximal. See, for example, the discussion of general enforcement in section 17.

First observe that important features of actual public law enforcement are congruent, at least in a broad sense, with what is theoretically desirable. Public enforcement is often characterized by low probabilities of detection. This is true for many criminal acts, and also is frequently the case in other spheres of public enforcement, such as traffic control and tax collection.[30] That probabilities of detection are low undoubtedly reflects the cost of raising the probability, a central factor in our discussion.

Corresponding to the low probabilities of detection are relatively high sanctions, often exceeding harm. For example, it seems that the sentence for theft typically outweighs the harm from that act, that the penalty for double parking frequently surpasses the resulting congestion costs, and that the sanction for tax evasion tends to exceed the social losses thereby created. Sanctions that are in excess of harm are needed for proper deterrence when the probabilities of enforcement are less than one, as they are in these examples.

The theory of optimal public enforcement of law also helps to explain why society uses the sanction of imprisonment when it does—for the category of harmful acts labeled criminal, notably, for theft, robbery, rape, murder, and so forth.[31] Because such acts cause substantial harm, yet often are detected with a low probability, the magnitudes of desirable penalties are high. If these penalties were solely monetary, they often would exceed the assets of the individuals who commit the acts, for individuals who commit crimes tend to have very low assets.[32] Imprisonment sanctions, therefore, usually will be required to maintain an adequate level of deterrence of acts classified as criminal.[33]

Note, too, that the standard of liability when imprisonment sanctions are imposed is generally fault-based—imprisonment is premised on the nature of the wrongful act, not merely on the fact that harm was done. This is socially desirable because, as we stressed, fault-based liability reduces the use of socially costly sanctions.

Although actual public enforcement is consistent in many respects with the theory of optimal enforcement, actual enforcement also appears to deviate in various ways from what is theoretically desirable. We note two discrepancies of general importance. First, substantial enforcement costs could be saved without sacrificing deterrence by reducing

[30] U.S. Department of Justice (1997b, p. 205, table 25) indicates, for example, that the likelihood of arrest was 13.8% for burglary, 14.0% for automobile theft, and 16.5% for arson. Kenkel (1993, p. 145) estimates that the probability of arrest for drunk driving is "only about .003." And according to Andreoni et al. (1998, p. 820), the audit rate for individual tax returns was 1.7 percent in 1995.

[31] See generally Posner (1985, pp. 1201–1205), Shavell (1985, pp. 1236–1241), and Shavell (2004, pp. 543–568).

[32] For example, in U.S. Department of Justice (1988, p. 35) it is reported that "the average inmate was at the poverty level before entering jail" and in U.S. Department of Justice (1998, p. 4) it is stated that almost half of jail inmates reported incomes of less than $600 a month in the month before their most recent arrest.

[33] The use of imprisonment sanctions also makes sense in view of their incapacitative function: crimes cause substantial harm and may be difficult or expensive to deter (for the reason we just emphasized, as well as others), so that it often will be desirable to incapacitate individuals who have committed them. See section 25.

enforcement effort and simultaneously raising fines. This is possible in many enforcement contexts because fines are presently very low relative to the assets of violators. For example, fines for most parking violations are less than $50, penalties for underpayment of income taxes are typically on the order of 20% of the amount not paid, and fines for corporate violations of health and safety regulations are frequently minuscule in relation to corporate assets. In such areas of enforcement, therefore, fines could readily be, say, doubled and enforcement costs reduced significantly, while maintaining deterrence at present levels.

Not only can present levels of deterrence be achieved more cheaply, it also seems that these levels are often too low. This is a reasonable supposition given the limited use of fines that we just noted and the low probabilities of their application. For example, the probability of a tax audit is approximately 2%; when combined with the modest penalties for underpayment, one would predict substantial tax avoidance.[34] Evidence also suggests that the expected fine for driving while intoxicated is on the order of one-quarter of the expected harm caused by such behavior,[35] and that monetary sanctions imposed on corporations equal on average only thirty-three percent of the harms caused.[36] Given the ample opportunities that exist for augmenting penalties, as well as the possible desirability of increasing enforcement effort, society probably should raise levels of deterrence in many areas of enforcement.

Part D: Extensions of the basic theory

This concludes the presentation of the basic theory of public enforcement of law. We now turn to various extensions and refinements of the basic analysis.

12. Accidental harms

In our analysis above, we implicitly assumed that the acts that individuals commit result in harm with certainty. In many circumstances, of course, acts result in harm only with

[34] In 1995 the audit rate for individual returns was 1.7 percent, as noted above, and the civil penalty for underpayment of taxes ordinarily is calculated as 20 percent of the underpayment that results from wrongful conduct (such as substantially misstating a valuation). See Andreoni et al. (1998, p. 820). Thus, for every dollar of underpayment, the expected payment, including the underpayment and the civil penalty, is only approximately $0.02 (= .017 × $1.20).

[35] See Kenkel (1993, p. 145). The expected fine is $12.82 and the expected harm is $47.77 (both in 1986 dollars). While the latter number may seem low, keep in mind that it is the product of the probability that a harm will occur as a result of drunken driving, and the level of harm if harm does occur. (To properly determine whether dangerous driving is underdeterred, one also would have to take into account the threat of liability from private suits brought by accident victims. But the deterrent effect of such suits will be dulled to the extent that drivers do not have sufficient assets to pay for the harms suffered by accident victims, or have liability insurance and therefore only partially bear the financial consequences of a lawsuit.)

[36] See Cohen (1989, pp. 617–618, 658). Cohen notes, however, that he did not take into account other sanctions imposed on corporate criminals, including restitution, civil penalties, and private tort suits.

a probability. A driver that speeds only creates a likelihood of a collision; or a firm that stores toxic chemicals in a substandard tank only creates a probability of a harmful spill.

Essentially all of the results in the basic analysis carry over in a straightforward way when harms are accidental. If individuals are risk neutral, sanctions are monetary, and the expected sanction equals harm whenever harm turns out to occur, then induced behavior will still be socially optimal; further, the optimal magnitude of sanctions is maximal if individuals are risk neutral because this allows enforcement costs to be saved, but is not necessarily maximal if individuals are risk averse, and so forth. Our general conclusions in the basic analysis can thus be interpreted to apply both when harms occur for sure and when harms occur accidentally.

There is, however, a new issue that arises when harm is uncertain: a sanction can be imposed either on the basis of the commission of a dangerous *act* that increases the chance of harm—storing chemicals in a substandard tank—or on the basis of the *actual occurrence of harm*—only if the tank ruptures and results in a spill. In principle, either approach can be employed to achieve optimal deterrence. To illustrate, suppose that the substandard tank has a 10% chance of rupturing, in which case the harm would be $10 million; the expected harm from using the tank therefore is $1 million. If individuals are risk neutral and sanctions are imposed only when harm occurs, deterrence will be optimal if, as usual, the sanction equals the harm of $10 million (assuming for simplicity that individuals are always found liable). Alternatively, if sanctions are imposed on the basis of the dangerous act of using the substandard tank, deterrence will be optimal if the owner of the tank faces a sanction equal to the expected harm due to his use of the substandard tank, $1 million.

Several factors are relevant to the choice between act-based and harm-based sanctions. First, act-based sanctions need not be as high to accomplish a given level of deterrence, and thus offer an advantage over harm-based sanctions because of limitations in parties' assets. In the example in the preceding paragraph, the owner of the storage tank might be able to pay the $1 million required if sanctions are act-based but not the $10 million required if sanctions are harm-based. Second, and closely related, because act-based sanctions need not be as high to accomplish deterrence, they tend to be preferable to harm-based sanctions when parties are risk averse. Third, act-based sanctions and harm-based sanctions may differ in the ease with which they can be applied. In some circumstances, act-based sanctions may be simpler to impose (it might be less difficult to determine whether an oil shipper properly maintains its vessels' holding tanks than to detect whether one of the vessels leaked oil into the ocean); in other circumstances, harm-based sanctions may be more readily applied (a driver who causes an accident might be caught more easily than one who speeds but does not cause an accident). Fourth, it may be hard to calculate the expected harm due to an act, but relatively easy to ascertain the actual harm if it eventuates; if so, this constitutes an advantage of harm-based liability.[37]

[37] Act-based versus harm-based enforcement is discussed in Shavell (1993).

13. Precautions

In this section we consider a model in which harm is accidental, as in the previous section, and in which the probability of harm depends on the level of precautions taken by a potential injurer. Thus, the major difference from the basic model considered in earlier sections is that the act is continuously variable. The main results of the basic analysis carry over to the model of precautions. For simplicity, we focus on the case in which enforcement is certain. Let

$$x = \text{level of precautions taken by a potential injurer; and}$$

$$q(x) = \text{probability of harm given } x;\ q'(x) < 0;\ q''(x) > 0.$$

The usual social objective of maximizing social welfare now can be expressed as minimizing social costs, that is, minimizing the sum of the cost of precautions and the expected harm: $x + q(x)h$. Let $x^* > 0$ be the solution to this problem.

First consider strict liability when the sanction is a fine equal to the harm. Then an individual's problem is to minimize $x + q(x)h$, so he will choose x^* (and obviously would not if the fine did not equal harm).

Next consider fault-based liability when the standard corresponds to x^* and the sanction is a fine equal to harm. If an individual takes less precaution than x^*, he bears costs of $x + q(x)h$, while if he takes precaution equal to or greater than x^*, he bears cost of x. It is straightforward to show that he will exercise precautions equal to x^*.[38]

Thus, as in the basic theory, strict liability and fault-based liability result in the first-best outcome when sanctions are monetary and are applied for sure. Similar reasoning would demonstrate that all of the other primary conclusions in the basic theory would carry over to the model of precautions.

To illustrate, reconsider the case when individuals are risk neutral, liability is strict, and the sanction is a fine. Social costs are $x + q(x)h + e$. The level of precautions x is determined by the individual minimizing $x + q(x)p(e)f$. Again, the optimal fine must be the individual's wealth w (otherwise, f could be raised and e lowered without affecting deterrence) and the optimal p is such that $pw < h$.[39] Therefore, at the optimum, the level of precautions is less than the first-best level.

[38] Conditional on choosing $x \geq x^*$, his best choice clearly is x^*, as that minimizes his expense, which is x^*. If he chooses $x < x^*$, his expense is $x + q(x)h$, which exceeds $x^* + q(x^*)h > x^*$. Hence, x^* is strictly optimal for him.

[39] Let $x(e)$ be the x determined by the individual's optimization problem, given $f = w$ and enforcement expenditures e. The individual minimizes $x + q(x)p(e)w$, with the resulting first-order condition $1 + q'p(e)w = 0$. The social problem is to minimize $x(e) + q(x(e))h + e$. The first-order condition can be written as $x'(1 + q'h) + 1 = 0$. We know that $x' > 0$. Therefore, it must be that $1 + q'h < 0$. Solving for q' from the individual's first-order condition and substituting it into this expression implies that $pw < h$.

14. Activity level

We have been assuming that the sole decision that an individual makes is whether to act in a way that causes harm when engaging in some activity. In many contexts, however, an individual also makes a choice about his *activity level*—that is, not only does he choose whether to act in a harmful way while engaging in an activity, he also chooses whether to engage in that activity, or, more generally, at what level to do so. For example, in addition to deciding whether to comply with auto emissions controls (maintaining a catalytic converter), an individual also chooses how many miles to drive; the number of miles driven is the individual's level of activity. Similarly, not only does a firm decide whether to comply with workplace safety regulations, it also chooses its level of production; the output of the firm is its level of activity.

The socially optimal activity level is such that the individual's marginal utility from the activity just equals the marginal expected harm caused by the activity. Thus, the optimal number of miles driven is the level at which the marginal utility of driving an extra mile just equals the marginal expected harm per mile driven. The determination of the optimal level of activity presumes that individuals act optimally when engaging in the activity—for example, that they drive with appropriate care.

To illustrate this formally, let

$$r = \text{level of activity; and}$$

$$U(r) = \text{utility from activity level } r;\ U'(r) > 0;\ U''(r) < 0.$$

We suppose that an individual chooses how much precaution to take (see the previous section) while engaging in a harm-creating activity, with the level of harm being proportional to his level of activity.[40] Then social welfare is

$$U(r) - r[x + q(x)h]. \tag{20}$$

Note that the optimal level of precaution x^* minimizes $x + q(x)h$, and thus is as discussed in section 13. The optimal level of activity therefore is determined by

$$U'(r) = x^* + q(x^*)h; \tag{21}$$

that is, the marginal utility from the activity equals the social cost of the activity, which is the sum of the cost of precautions and the expected harm.

Will parties' choices about their activity levels and precautions be socially correct under the two standards for imposing sanctions? The answer is that under strict liability, their choices about both activity levels and precautions will be socially correct. This is clear since, assuming for simplicity that enforcement is certain and the sanction is a fine equal to harm, their objective also is to maximize (20). In particular, because they bear

[40] We are employing the model of precautions described in the previous section for convenience. It also would be possible to develop the points about activity level using the model from the basic analysis, in which the harm-producing action of an individual is not continuously variable.

a fine equal to harm, they choose the optimal level of precautions x^* when engaging in their activity. And since they incur the full social costs of precautions plus expected harm when engaging in their activity, they choose the optimal level of activity.

Under the fault-based standard, however, parties will participate in activities to a socially excessive extent. To explain, observe that parties choose the optimal level of precautions x^* in order to avoid the fine, as seen in section 13. Because parties choose this level of precautions, they will not be found liable for having violated the standard if harm occurs. Hence, their choice of activity level r is determined by maximizing $U(r) - rx^*$, with the corresponding first-order condition

$$U'(r) = x^*. \tag{22}$$

Comparing this to (21), it is evident that r exceeds r^*. The reason is that the private marginal cost of increasing participation in the activity is only the precaution cost x^*; it does not include the expected harm. Thus, for instance, if a person complies with auto emissions standards, he will not be concerned with the fact that the more he drives, the more pollution he causes (assuming that some pollution occurs even if one obeys emissions standards). Consequently, he will drive too much.

The implication of the preceding points in relation to firms is that under the strict sanctioning rule, the product price will reflect both the cost of precautions and the expected harm caused by production, so that the price will include the full social cost of production. Hence, the amount purchased, and thus the level of production, will be socially optimal. Under the fault-based rule, however, the product price will reflect the cost of precautions but not expected harm; thus, the amount sold, and the level of production, will be excessive.[41]

These conclusions about firms are of widespread applicability. Notably, safety regulations and other regulatory requirements are often framed as standards of care that have to be met, but which, if met, free the regulated party from penalties. Hence, regulations of this character are subject to the criticism that they lead to excessive levels of the regulated activity. Making firms strictly liable for harm would be superior to safety regulation with respect to inducing socially correct activity levels.

That parties choose an excessive level of activity under the fault-based rule—of which regulation is one variant—but not under the strict liability rule, constitutes a fundamental advantage of the latter rule. This advantage is stronger the greater is the harm from engaging in the activity (given that precautions are optimal when engaging in the activity). Thus, for activities for which expected harm is likely to be substantial, the disadvantage of the fault-based standard is significant.

More generally, the advantage of strict liability over fault-based liability applies to any dimension of behavior that affects expected harm but that is not included in the definition of fault. For example, suppose that pollution damage depends both on whether

[41] Our discussion here about activity-level considerations in the context of public enforcement closely parallels the analysis of activity-level issues in the context of tort liability. See generally Shavell (1980) and Polinsky (1980b).

a scrubber is installed as well as on the degree of care with which it is cleaned. Because the existence of a scrubber is easy to verify but its maintenance might not be, a fault-based sanctioning system might, of necessity, reflect only the first dimension of behavior. Consequently, the firm will have a socially inadequate incentive to clean the scrubber under a fault standard. This problem does not arise under a strict sanctioning system because the firm has to pay for harm regardless of its cause.

15. Errors

Errors of the two classic types can occur in public enforcement of law. First, an individual who should be found liable might mistakenly not be found liable—what we will refer to as "mistaken acquittal." Second, an individual who should not be found liable might mistakenly be found liable—"mistaken conviction." For an individual who has been detected, let

$\varepsilon_A =$ the probability of mistaken acquittal; and

$\varepsilon_C =$ the probability of mistaken conviction.

For example, suppose police randomly monitor drivers by stopping them and administering a blood-alcohol test. The test might understate the amount of alcohol in the driver's blood and result in mistaken acquittal, or overstate the amount and lead to mistaken conviction.

We initially consider the effect of mistake in the basic model of enforcement, assuming that the sanctioning standard is strict, the sanction is a fine, and individuals are risk neutral. Given the probability of detection p and the chances of mistaken acquittal and conviction, an individual will commit the wrongful act if and only if his gain net of his expected fine if he does commit it exceeds what he bears if he does not commit it:

$$g - p(1 - \varepsilon_A)f > -p\varepsilon_C f, \tag{23}$$

or, equivalently, if and only if[42]

$$g > (1 - \varepsilon_A - \varepsilon_C)pf. \tag{24}$$

Note initially that both types of error reduce deterrence: the right-hand side of (24) is declining in both ε_A and ε_C. Mistaken acquittal diminishes deterrence because it lowers the expected fine if an individual violates the law. Mistaken conviction also lowers deterrence because it reduces the difference between the expected fine from violating the law and not violating it. In other words, the greater is ε_C, the smaller the increase in the expected fine if one violates the law, making a violation less costly to the individual.[43]

[42] We assume that $1 - \varepsilon_A - \varepsilon_C > 0$, so that the probability that a guilty person will be found liable, $1 - \varepsilon_A$, exceeds the probability that an innocent person will be found liable, ε_C.

[43] This point was first emphasized by Png (1986).

Because mistakes dilute deterrence, they tend to reduce social welfare. Specifically, to achieve any level of deterrence, it may be necessary to raise the probability of detection or the magnitude of a costly sanction to offset the effect of errors.

Now consider the optimal choice of the fine. If the probability of detection is assumed to be fixed, the dilution in deterrence caused by errors requires a higher fine to restore deterrence, so the optimal fine is higher.[44] If both the probability and the fine are policy instruments, the optimal fine remains maximal despite mistakes. The explanation is essentially that given previously: If the fine f were less than maximal, then f could be raised and the probability p lowered so as to keep deterrence constant, but saving enforcement costs.

If individuals are risk averse, however, the possibility of mistakes does affect the optimal fine. As we emphasized in section 7.2, the optimal fine generally is less than maximal when individuals are risk averse—lowering the fine reduces the bearing of risk. Introducing the possibility of mistakes may increase the desirability of lowering the fine because, due to mistaken conviction, individuals who do not violate the law are subject to the risk of having to pay a fine. Indeed, because the number of persons who do not violate the law often would far exceed the number who do, the desire to avoid imposing risk on the former group can lead to a substantial reduction in the optimal fine.[45]

The possibility of mistakes generally affects the optimal probability of detection. On one hand, the deterrence-diluting effects of mistakes means, as we noted, that a higher probability of detection may be needed to achieve any given level of deterrence; this effect tends to raise the optimal expenditure on enforcement. On the other hand, because mistakes reduce the productivity of enforcement expenditures by a factor of $1 - \varepsilon_A - \varepsilon_C$ (see (24)), the cost of achieving a given level of deterrence is higher; this effect tends to reduce the optimal expenditure on enforcement. Either of these effects could dominate and lead to an optimal probability of detection that is higher or lower than in the absence of mistakes.

Next, consider imprisonment and mistake. As in the case of fines, mistakes of both type dilute the deterrent effect of imprisonment. Additionally, as in the case without mistakes, the optimal imprisonment term is maximal if individuals are risk neutral or risk averse in imprisonment, but is generally not maximal if they are risk preferring in imprisonment.[46]

[44] Specifically, to achieve first-best behavior, it must be that $(1 - \varepsilon_A - \varepsilon_C)pf = h$, which implies that f must be higher the greater are either of the errors, ε_A or ε_C.

[45] Building on Polinsky and Shavell (1979), Block and Sidak (1980, pp. 1135–1139) emphasize the desirability of lowering sanctions on risk-averse individuals because of mistakes.

[46] That the optimal term remains maximal if individuals are risk neutral or risk averse might seem surprising because one might expect that the chance of mistaken conviction would result in a lower optimal term. But the usual argument still applies: If the term were not maximal, it could be raised and the probability of detection could be lowered at least proportionally without sacrificing deterrence. Hence, the aggregate amount of jail time served by individuals who do not commit the harmful act would remain the same or fall, and enforcement expenditures would decline.

The possibility of mistakes also affects individuals' decisions regarding their participation in activities. Everything else equal, mistaken acquittals lead to increased engagement in the activity, and mistaken convictions result in decreased engagement. The net effect could be a socially excessive or inadequate activity level.[47]

We have not yet commented on fault-based liability and mistake. In this context, an important implication of mistake is that some individuals will bear sanctions even if they comply with the fault standard. Consequently, both types of error reduce the incentive to comply with the fault standard in the basic model, for essentially the same reasons as under strict liability.

However, in the model with variable precautions, there is a possibility that error will lead to an excessive level of precautions—above the optimal level. Assume that the actual level of precautions x is observed with an error ε. If the observed level of precautions, $x + \varepsilon$, exceeds the standard $\bar{x}$, the person will not be liable because he will not be found to be at fault. But if $x + \varepsilon$ is less than $\bar{x}$, then the person will be liable. The person will be found mistakenly liable if x is greater than or equal to $\bar{x}$ but $x + \varepsilon$ is less than $\bar{x}$; the person will be mistakenly acquitted if x is less than $\bar{x}$ but $x + \varepsilon$ equals or exceeds $\bar{x}$. Note that the probabilities of the two types of error depend on the person's choice of x. By choosing an x above $\bar{x}$, the person can reduce the risk of mistaken conviction, and it can be shown that he will be led to choose such an excessive x under fairly general conditions. In other words, individuals will often have a motive to take excessive precautions in order to reduce the chance of erroneously being found at fault.[48]

Errors also influence individuals' participation in the activity under fault-based liability, similar to the effects of errors under strict liability. The main difference is that there is a general tendency for individuals to participate in the activity to an excessive extent under fault-based liability, for the reason explained in the previous section. Only if the chance of mistaken conviction is sufficiently high would this conclusion be reversed.

Finally, observe that the probabilities of error can be influenced by policy choices. For example, prosecutorial resources can be increased in order to reduce the probability of mistaken acquittal, or the standard of proof can be raised to reduce the chance of mistaken conviction (although this presumably increases the likelihood of mistaken acquittal). Because the reduction of both types of error increases deterrence, expenditures made to reduce errors may be socially beneficial.[49]

[47] Recall, too, that if the probability of detection is fixed, the fine needs to be raised to offset the deterrence-diluting effects of mistakes. Raising the fine to this extent, however, leads to an inadequate incentive to engage in the activity. This problem, in turn, can be remedied by use of an appropriate subsidy for participating in the activity. For the details behind this point, see Png (1986).

[48] This point was first emphasized by Craswell and Calfee (1986).

[49] On the value of accuracy in adjudication, see Kaplow and Shavell (1994a).

16. Costs of imposing fines

We inquire in this section about the implications of costs borne by enforcement authorities in imposing fines.[50] Our principal observation is that such costs should raise the level of the fine.

To elaborate, suppose that the probability of detection is fixed at p, that liability is strict, and that individuals are risk neutral. If fines are costless to impose, the optimal fine is h/p, the harm divided by the probability of detection (see (14)). Now suppose that the enforcement authority bears a cost each time a fine is imposed; let

$k =$ cost of imposing a fine on an offender.

It is easy to verify that the optimal fine then is

$$f^* = h/p + k; \tag{25}$$

the cost k should be added to the fine that would otherwise be desirable. The explanation is that, if an individual commits a harmful act, he causes society to bear not only the immediate harm h, but also, with probability p, the cost k of imposing the fine—that is, his act results in an expected total social cost of $h + pk$. If the fine is set according to (25), the individual's expected fine is $h + pk$, which leads him to engage in the harmful act if and only if his gain exceeds the expected total social cost of the act.

There may be other costs associated with the imposition of fines. In particular, suppose that detection is followed by a costly second stage during which the state investigates and prosecutes an individual, and at the end of which a fine is imposed only with a probability. Let

$s =$ cost of the investigation-prosecution stage; and

$q =$ probability of a fine being imposed after the investigation-prosecution stage.

Hence, the probability that an individual will have to pay a fine is pq and the expected costs of imposing a fine, including the expected investigation-prosecution cost, become $ps + pqk$.

It is readily shown that the optimal fine now is

$$f^* = h/pq + s/q + k. \tag{26}$$

This formula illustrates a general principle: the optimal fine equals the costs incurred by society as a result of the harmful act divided by the probability—at the time that each component of cost is incurred—that the individual will have to pay the fine. Thus, h is divided by pq because, when the harm occurs, the probability of having to pay the fine is pq; and s is divided by q because, when the investigation-prosecution costs are

[50] We have already discussed the cost of imposing imprisonment sanctions—specifically, the cost to the state per unit of the imprisonment term, c.

incurred, the probability of having to pay the fine is q. If the fine is computed according to this principle, the expected fine will equal the expected social costs due to an individual committing a harmful act, including the harm caused and the expected sanctioning costs—that is, $h + ps + pqk$.

Note that under fault-based liability, the costs of imposing fines is significantly lower, if not zero. This is because, if individuals comply with the fault standard, they do not bear sanctions, in which case there are no costs associated with imposing sanctions. However, if individuals are found at fault (say because of errors), the fines imposed on them also should reflect the costs of imposing fines.

Finally, observe that not only does the state incur costs when fines are imposed, so do the individuals who pay the fines (such as legal defense expenses). The costs borne by individuals, however, do not affect the formula for the optimal fine. Individuals properly take these costs into account because they bear them directly.[51]

17. General enforcement

In many settings, enforcement may be said to be *general* in the sense that several different types of violations may be detected by an enforcement agent's activity. For example, a police officer waiting along the side of a road may notice a driver who litters as well as a driver who goes through a red light or who speeds; or a tax auditor may detect a variety of infractions when he examines a tax return. To analyze this type of situation, suppose that a single probability of detection applies to all harmful acts, regardless of the magnitude of the harm.[52] (The contrasting assumption is that enforcement is *specific*, meaning that the probability is chosen independently for each type of harmful act.)

The main point that we want to make is that when enforcement is general, the optimal sanction rises with the severity of the harm and is maximal only for relatively high harms. To see this, assume that liability is strict, the sanction is a fine, and individuals are risk neutral. Let $f(h)$ be the fine given harm h. Then, for any general probability of detection p, the optimal fine schedule is

$$f^*(h) = h/p, \tag{27}$$

provided that h/p does not exceed the maximum feasible fine (say individuals' wealth level w); if h/p is not feasible, the optimal fine is maximal. This schedule is obviously

[51] The points developed in this section were first presented in Polinsky and Shavell (1992), although early writers on enforcement theory—including Becker (1968, p. 192) and Stigler (1970, p. 533)—recognized that sanctions should reflect enforcement costs.

[52] It will be clear that the main point developed in this section does not depend on the assumption that the same probability applies to all acts. The only requirement is that the probabilities for different acts are linked, all a function of the same enforcement expenditure.

optimal given p because it implies that the expected fine equals harm, thereby inducing first-best behavior, whenever that is possible.

The question remains whether it would be desirable to lower p and raise fines to the maximal level for the low-harm acts for which $f^*(h)$ is less than maximal. But if p is reduced for the relatively low-harm acts (and the fine raised for them), then p—being general—is also reduced for the high-harm acts for which the fine is already maximal, resulting in lower deterrence of these acts. The decline in deterrence of high-harm acts may cause a greater social loss than the savings in enforcement costs from lowering p. To express this point differently, p must be sufficiently high to avoid significant underdeterrence of high-harm acts (for which fines are maximal). But since this p also applies to less harmful acts, the fines for them do not need to be maximal in order to deter them appropriately.[53]

The result that, when enforcement is general, sanctions should rise with the severity of harm up to a maximum also holds if the sanction is imprisonment and if liability is fault-based. The underlying reasoning is the same as that given above.[54]

18. Marginal deterrence

In many circumstances, an individual may consider which of *several* harmful acts to commit, for example, whether to release only a small amount of a pollutant into a river or a large amount, or whether only to kidnap a person or also to kill him. In such contexts, the threat of sanctions plays a role in addition to the usual one of deterring individuals from committing harmful acts: for individuals who are not deterred, expected sanctions influence *which* harmful acts individuals choose to commit. Notably, such individuals will have a reason to commit less harmful rather than more harmful acts if expected sanctions rise with harm. Deterring a more harmful act by having its expected sanction exceed that for a less harmful act is sometimes referred to as *marginal deterrence*.[55]

Other things being equal, it is socially desirable that enforcement policy creates marginal deterrence, so that those who are not deterred from committing harmful acts have a reason to moderate the amount of harm that they cause. This suggests that sanctions should rise with the magnitude of harm and, therefore, that most sanctions should be less than maximal. However, promoting marginal deterrence may conflict with achieving deterrence generally: for the schedule of sanctions to rise steeply enough to accomplish

[53] Note that if p could be varied independently for a low-harm act and for a high-harm act—that is, if enforcement is specific rather than general—then it would be desirable to lower p and raise the fine for a low-harm act if the fine for it were less than maximal.

[54] The basic point of this section was first made by Shavell (1991b); see also Mookherjee and Png (1992) for a closely related analysis.

[55] The notion of marginal deterrence was remarked upon in some of the earliest writing on enforcement; see Beccaria (1767, p. 32) and Bentham (1789, p. 171). The term *marginal deterrence* apparently was first used by Stigler (1970).

marginal deterrence, sanctions for less harmful acts might have to be so low that individuals are inadequately deterred from committing these acts.[56]

To illustrate the implications of marginal deterrence, consider the following example in which sanctions are monetary and liability is strict. Suppose that there are two harmful acts, with harms h_1 and h_2, where $h_1 < h_2$, that the probability of detection p is the same for both acts, and that individuals have the same level of wealth w. We will first consider a one-act model in which there can be no marginal deterrence because each individual can commit only one type of harmful act. We will then compare the results in this case to a two-act model in which each individual can commit either of two harmful acts.

In the one-act model, suppose some individuals have the opportunity to commit an act causing harm of h_1 and other individuals have the opportunity to commit an act causing harm of h_2. It is optimal to set the fine for the high-harm act equal to w, for otherwise it would be possible to raise the fine for both acts and lower the probability of detection p without affecting deterrence, but saving enforcement costs. It also follows that the optimal p is such that pw is less than h_2, that is, there is some underdeterrence of the high-harm act. The reason is that if $pw = h_2$, there would be no first-order loss of social welfare in terms of gains and harm if p is lowered (since marginal individuals are those for whom $g = h_2$), but enforcement costs would be saved. Given the common probability p, the fine f_1 for the lesser offense then can be set such that $pf_1 = h_1$ (assuming such a fine is feasible), achieving first-best deterrence of this offense.

In the two-act model, each individual can commit an act causing harm of h_1 or an act causing harm of h_2. Again, it is optimal to set the fine for the high-harm act equal to w and for there to be underdeterrence of the high-harm act. Now, however, f_1 should be such that pf_1 is less than h_1, instead of equaling h_1. The essential reason for this result is that the reduction in f_1 from h_1/p leads some offenders to commit the act causing harm h_1 instead of the act causing higher harm h_2. This can be shown to raise social welfare even though the reduction in f_1 leads some individuals to commit the low-harm act who otherwise would not have committed either harmful act.[57] In other words, achieving marginal deterrence by reducing the expected fine for the low-harm act raises social welfare.

Two additional observations should be made about marginal deterrence. First, marginal deterrence can be promoted by increasing the probability of detection as well as the magnitude of sanctions for acts that cause greater harm. For example, kidnappers can be deterred more from killing their victims if greater police resources are devoted

[56] For formal treatments of marginal deterrence, see Shavell (1992), Wilde (1992), and Mookherjee and Png (1994).

[57] This conclusion essentially follows from two observations. First, because pw is less than h_2, some individuals who had been committing the high-harm act were causing a net loss of social welfare (their gain was less than h_2). Second, as f_1 is lowered marginally from h_1/p, individuals who are induced to commit the act causing harm h_1 (who either would have committed the high-harm act or not committed any harmful act) cause no net loss of social welfare (their gain equaled h_1).

to apprehending kidnappers who murder their victims than to those who do not. (Note, though, that in circumstances in which enforcement is general, the probability of detection cannot be independently altered for acts that cause different degrees of harm.) Second, marginal deterrence is naturally accomplished if the expected sanction equals harm for all levels of harm; for if a person is paying for harm done, he will have to pay appropriately more if he does greater harm. Thus, for instance, if a polluter's expected fine would rise from \$100 to \$500 if he dumps five gallons instead of one gallon of waste into a lake, where each gallon causes \$100 of harm, his marginal incentives to pollute will be correct.[58]

19. Principal–agent relationship

Although we have assumed that an offender is an independent single actor, in fact the offender is often an agent of a principal. For example, the agent could be an employee of a firm or a subcontractor of a contractor. The enforcement problem is now how to maximize social welfare by choosing enforcement effort and the sanctions to be imposed on principals and agents. This maximization is carried out under the assumption that a principal chooses a contract with his agent that maximizes the principal's expected utility subject to two constraints: that the agent receive his reservation level of expected utility; and that the agent maximizes his own expected utility.

When harm is caused by an agent, many of our conclusions from the basic analysis carry over if the sanction is imposed on the principal. For example, given the probability of detection p, it is optimal for a risk-neutral principal to face a fine of h/p. Then the expected fine is equal to harm done. Consequently, the principal will behave socially optimally in controlling his agents, and in particular will contract with them and monitor them in ways that will give the agents socially appropriate incentives to reduce harm.[59]

A question about enforcement that arises when there are principals and agents is how to allocate financial sanctions between them. First observe that the particular allocation of sanctions may not matter when, as would be the natural presumption, the principal

[58] As we discussed in section 7.1, however, it generally is desirable for society to tolerate some underdeterrence in order to save enforcement costs, in which case expected sanctions will be less than harm. Then, consideration of marginal deterrence alters the structure of sanctions that would otherwise be best, as the comparison of the one-act model to the two-act model in this section showed.

[59] There is relatively little literature on the question of optimal enforcement when wrongdoers are agents of principals. Newman and Wright (1990) study the optimal monetary sanction to impose on a risk-neutral principal when liability is strict and is imposed for sure; they show that it equals harm. Polinsky and Shavell (1993) demonstrate the potential desirability of imposing criminal sanctions on an employee who causes harm, even when the employer is capable of paying for the harm. Arlen (1994) examines the effect of sanctions on corporations' incentives to monitor their employees, and she emphasizes the possibility that corporations may have perverse incentives not to monitor if they would become liable as a result of their discovering and reporting employee violations. Also, Shavell (1997b) finds that optimal sanctions on corporations could be above or below harm when employee assets are less than harm.

and the agent can reallocate sanctions through their own contract. For example, if the agent finds that he faces a large fine but is more risk averse than the principal, the principal can assume it; conversely, if the fine would be imposed on the principal, he can bear that risk and not impose an internal sanction on the agent. Thus, the post-contract penalties that the agent suffers may not be affected by the particular division of sanctions initially selected by the enforcement authority.

The allocation of monetary sanctions between principals and agents would matter, however, if some allocations allow the pair to reduce their total burden. An important example is when a fine is imposed only on the agent and he is unable to pay it because his assets are less than the fine.[60] Then he and the principal (who often would have higher assets) would jointly escape part of the fine, diluting deterrence. The fine therefore should be imposed on the principal rather than on the agent (or at least the part of the fine that the agent cannot pay).

A closely related point is that the imposition of imprisonment sanctions on agents may be desirable when their assets are less than the optimal fine, even if the principal's assets are sufficient to pay the fine. The fact that an agent's assets are limited means that the principal may be unable to control him adequately through the use of contractually-determined penalties, which can only be monetary. For example, a firm may not be able, despite the threat of salary reduction or dismissal, to induce its employees never to rig bids. In such circumstances, it may be socially valuable to use the threat of personal criminal liability and a jail sentence to better control agents' misconduct.[61]

20. Settlements

We have thus far assumed that when an individual who should be found liable is discovered, he will be sanctioned in some automatic fashion. In practice, however, an individual must be found liable in a trial, and before this occurs, it is common for an individual to settle in lieu of trial. (In the criminal context, the settlement usually takes the form of a *plea bargain*, an agreement in which the individual pleads guilty to a reduced charge.) Given the prevalence of settlements, it is important to consider how they affect deterrence and the optimal system of public enforcement, and whether settlements are socially desirable.

A general reason why a wrongdoer who has been caught might prefer an out-of-court settlement to a trial is that a settlement saves him time and/or money. Public enforcers presumably would prefer settlements for this reason as well. To amplify, consider a risk-neutral detected individual subject to a fine f and let

$r =$ probability of conviction;

[60] See Sykes (1981) and Kornhauser (1982).

[61] This point is discussed by Segerson and Tietenberg (1992) and emphasized by Polinsky and Shavell (1993).

c_I = individual's litigation costs; and

c_P = prosecutor's litigation costs.

Assume that the parties agree on the probability of conviction. If the case goes to trial, the expected cost to the individual is $rf + c_I$, and the expected gain to the prosecutor, assuming his goal is to maximize expected penalties imposed net of his litigation costs, is $rf - c_P$. Thus, any settlement resulting in a fine between $rf - c_P$ and $rf + c_I$ would make both parties better off because the cost of litigation would be avoided. The same point would apply if the individual were subject to a jail sentence rather than a fine. Note, however, that if the parties disagree about the probability of conviction, a settlement might not occur—specifically, if the individual is relatively optimistic and believes that the probability of his being convicted is sufficiently less than the prosecutor thinks it is.

A second benefit of a settlement is that it eliminates the risks inherent in the trial outcome, a benefit to parties who are averse to such risks.[62]

The preceding advantages of settlement to the parties suggest that settlement is socially valuable, but the effect of settlement on deterrence is a complicating factor. Specifically, settlements dilute deterrence: for if individuals desire to settle, it must be because the expected disutility of sanctions is lowered for them. However, because settlements reflect the sanctions that would be imposed at trial, the state may be able to offset this settlement-related reduction in deterrence by increasing the level of sanctions. If so, settlements need not compromise the overall level of deterrence.[63]

Settlements may have other socially undesirable consequences. First, they may result in sanctions that are not as well tailored to harmful acts as would be true of court-determined sanctions. Second, settlements hinder the amplification and development of the law through the setting of precedents. Third, settlements also sometimes allow individuals to keep aspects of their behavior secret, which can reduce deterrence. Fourth, settlements for prison terms can result in increases in public expenditures on jail if individuals are risk averse in imprisonment.[64] A prosecutor whose goal is to maximize social welfare, as opposed to maximizing the expected sanction less prosecution costs,

[62] These benefits of settlement are well-recognized in the economic literature on civil litigation; see the surveys by Cooter and Rubinfeld (1989) and Spier (2007). For early discussions of settlement in the context of public enforcement, see Landes (1971) and Grossman and Katz (1983), and more recently, see, for example, Reinganum (1988), Polinsky and Rubinfeld (1989), Kobayashi and Lott (1992), and Miceli (1996).

[63] The deterrence-diluting effects of settlement and other aspects of the social desirability of settlement have been discussed in the private litigation context by Polinsky and Rubinfeld (1988), Shavell (1997a), and Spier (1997). A related discussion in the public enforcement context appears in Polinsky and Rubinfeld (1989).

[64] For example, suppose a defendant faces a 50% chance of a 5 year sentence and a 50% chance of a 15 year sentence, with an expected sentence of 10 years. If he is risk averse, he will strictly prefer a certain sentence of 10 years. This implies that if prosecutors want to maintain deterrence, they must demand a settlement of more than 10 years, say 12 years, which increases the cost of imprisonment.

presumably would take these additional factors into account and sometimes refuse to settle even though the settlement saves litigation costs and avoids risk.[65]

21. Self-reporting

We have assumed that individuals are subject to sanctions only if they are detected by an enforcement agent, but in fact parties sometimes disclose their own violations to enforcement authorities. For example, firms often report violations of environmental and safety regulations, individuals usually notify police of their involvement in traffic accidents, and even criminals occasionally turn themselves in.

Self-reporting can be induced by lowering the sanction for individuals who disclose their own infractions. To avoid significantly reducing deterrence, however, the reward for self-reporting can be made relatively small. For example, suppose that the fine if an individual does not self report is $1,000 and that the probability of detection is 10%, so the expected fine is $100. If the fine if one self-reports is $99, individuals will self-report but deterrence will barely be reduced.

To express this formally, assume for simplicity that individuals are risk neutral, and suppose that if an individual commits a violation and does not self-report, his expected fine is *pf*. Let

$$f' = \text{fine if a violator self-reports,}$$

and set

$$f' = pf - \varepsilon, \tag{28}$$

where $\varepsilon > 0$ is small. A violator will therefore want to self-report because f' is less than pf, but the deterrent effect of the sanction will approximate that if he did not self-report.

Given that self-reporting can be induced essentially without compromising deterrence, why is self-reporting socially advantageous? One reason is that self-reporting lowers enforcement costs because the enforcement authority does not have to identify and prove who the violator was. Environmental enforcers do not need to spend as much effort trying to detect pollution and establishing its source if firms that pollute report that fact.[66] Second, self-reporting reduces risk, and thus is advantageous if individuals are risk averse. Drivers bear less risk because they know that if they cause an accident, they

[65] The question of what prosecutors maximize has received almost no attention from law and economics scholars, although two exceptions are Miceli (1996) and Glaeser, Kessler, and Piehl (2000).

[66] In some contexts, however, self-reporting will not save enforcement costs. For example, suppose that a police officer waits by the side of a road to spot speeders. Then, were a driver to report that he had sped, this would not reduce policing costs, presuming that the officer still needs to be stationed at the roadside to watch for other speeders. Usually, though, there would be some cost savings as a result of self-reporting (for example, the police officer would not have to chase as many speeders).

will be led to report this to the police and suffer a lower and certain sanction (of approximately pf), rather than face a substantially higher sanction imposed only with some probability. Third, self-reporting may allow harm to be mitigated. Early identification of a toxic leak, for example, will facilitate its containment and clean-up.[67]

22. Repeat offenders

In practice, the law often sanctions repeat offenders more severely than first-time offenders. For example, under the U.S. Sentencing Commission's sentencing guidelines for Federal crimes, both imprisonment terms and criminal fines are enhanced if an offender has a prior record. Civil money penalties also sometimes depend on whether the offender has a record of prior offenses. We explain here why such policies may be socially desirable.

Note first that sanctioning repeat offenders more severely cannot be socially advantageous if deterrence always induces first-best behavior. If the sanction for polluting and causing a \$1,000 harm is \$1,000, then any person who pollutes and pays \$1,000 is a person whose gain from polluting (say the savings from not installing pollution control equipment) must have exceeded \$1,000. Social welfare therefore is higher as a result of his polluting. If such an individual polluted and was sanctioned in the past, that only means that it was socially desirable for him to have polluted previously. Raising the sanction because of his having a record of prior convictions would overdeter him now.

Accordingly, only if deterrence is inadequate is it possibly desirable to condition sanctions on offense history to increase deterrence. But deterrence often will be inadequate because, as we emphasized in section 7.1, it will usually be worthwhile for the state to tolerate some underdeterrence in order to reduce enforcement expenses.

Given that there is underdeterrence, making sanctions depend on offense history may be beneficial for two reasons. First, the use of offense history may create an additional incentive not to violate the law: if detection of a violation implies not only an immediate sanction, but also a higher sanction for a future violation, an individual will be deterred more from committing a violation presently.[68] Second, making sanctions depend on

[67] The basic theory of self-reporting in public enforcement is developed in Kaplow and Shavell (1994b); see also Malik (1993) and Innes (1999). Related literature concerns the reporting of income by individuals to tax authorities and the reporting of costs by regulated firms to regulatory authorities. See, for example, Andreoni, Erard, and Feinstein (1998) and Laffont and Tirole (1993).

[68] There is a subtlety in demonstrating the optimality of punishing repeat offenses more severely. Namely, if there is a problem of underdeterrence, one might wonder why it would not be optimal to raise the sanction to the maximum level for every offense (in which case repeat offenses would not be punished more severely). It must be shown that punishing all offenses maximally is inferior to punishing first offenses less than maximally and punishing repeat offenses more severely. See Polinsky and Shavell (1998) on the possible optimality of making sanctions depend on offense history because of the additional deterrence that such a policy creates. Miceli and Bucci (2005) supply a different reason for raising the fine for the second offense—that there is little additional social stigma associated with a second offense, so that a higher sanction is needed to maintain

offense history allows society to take advantage of information about the dangerousness of individuals and the need to deter them: individuals with offense histories may be more likely than average to commit future violations, which might make it desirable for purposes of deterrence to impose higher sanctions on them.[69]

There is also an incapacitation-based reason for making sanctions depend on offense history. Repeat offenders are more likely to have higher propensities to commit violations in the future and thus more likely to be worth incapacitating by imprisonment.[70]

23. Imperfect knowledge about the probability and magnitude of sanctions

Although we have made the simplifying assumption that individuals know the probability of detection and the magnitude of sanctions, it is obvious that individuals frequently have imperfect knowledge of these variables. They generally possess only subjective probability distributions of the probability of a sanction and its magnitude. They might not know the true probability of a sanction for several reasons: because the enforcement authority refrains from publishing information about the probability (perhaps hoping that individuals will believe it to be higher than it is); because the probability depends on factors that individuals do not fully understand (the probability of a tax audit, for example, is influenced by factors that are kept secret from taxpayers); and because probabilities might be difficult for individuals to assess.[71] Also, individuals may have incomplete knowledge of the true magnitudes of sanctions, particularly if sanctions are not fixed by law, but are to some degree discretionary.[72]

The implications of individuals' imperfect knowledge are straightforward to ascertain. First, to predict how individuals behave, what is relevant, of course, is not the actual probability and magnitude of a sanction, but perceptions of them.

Second, to determine the optimal probability and magnitude of sanctions, account must be taken of the relationship between the actual and the perceived values. To illustrate, suppose that the perceived probability is a single value $\hat{p}(p)$, where $\hat{p}$ is increasing in the true probability p, and, similarly, that the perceived fine is $\hat{f}(f)$, where $\hat{f}$ is increasing in the true fine f. Thus, if the probability of detection p is fixed, the optimal

deterrence. Emons (2003), however, raises the possibility that it may be optimal to lower the sanction for the second offense.

[69] Note that this reason for making sanctions depend on offense history is different from the first reason: the second reason involves the assumption that offenders are different and that the optimal sanction for some offenders is higher than for others; the first reason applies even if individuals are identical. On the second, information-based, reason for making sanctions depend on offense history, see Rubinstein (1979), Polinsky and Rubinfeld (1991), and Chu, Hu, and Huang (2000).

[70] See section 25 for a discussion of the incapacitation rationale for the use of imprisonment sanctions.

[71] On the problems that individuals have in evaluating and using probabilities, see, for example, Kahneman, Slovic, and Tversky (1982).

[72] In addition, individuals could have imperfect information about the prevailing standard of liability, not being sure whether it is strict or fault-based. This type of mistake, about a discrete issue, seems less likely to be significant than errors in assessing the probability and magnitude of sanctions.

fine is set such that $\hat{p}(p)\hat{f}(f) = h$, assuming such a fine is feasible. This might imply a higher or a lower fine than when perceptions are accurate.

Third, the result that the optimal fine should be maximal when individuals are risk neutral continues to hold. By raising the fine to the maximum, the perceived fine also will be maximal. The probability of detection then can be lowered, thereby saving enforcement costs without reducing deterrence.

Several other observations are worth making. One concerns lags in learning about changes in enforcement policy. For example, suppose that there is a delay of at least a year before individuals fully comprehend a change in the probability of enforcement. Then if enforcement resources are increased so as to make the probability, say, 15% rather than 10%, there might not be a significant increase in deterrence for some time, making such an investment less worthwhile.[73] Another observation involves the difficulty in learning about variations in enforcement policy when enforcement policy is described by a distribution. For instance, suppose that the sanction for some act, such as robbery, can vary (say from one month of jail time to 10 years), and that individuals' perceptions are quite rough, not based on true averages, but mainly on the possible range of sanctions. Then increasing the average sentence within this range might have very little effect on deterrence. The processes by which individuals formulate probabilities of sanctions and their magnitudes are important, therefore, to determining optimal deterrence policy.[74]

24. Corruption

In this section we examine the possible corruption of law enforcement agents, how corruption lessens deterrence and distorts participation in harm-creating activities, and what policies may be employed to combat corruption.[75] One form of corruption is bribery, in which a law enforcer accepts a payment in return for not reporting a violation (or for reducing the mandated sanction for the violation). For example, in consideration of a bribe payment, a police officer may overlook a speeding violation or a building inspector may ignore a code infraction. (For simplicity, we do not distinguish between

[73] Similarly, suppose that individuals treat all probabilities of enforcement that are low, say below 1%, as if they were probabilities of 1%, because it is not possible for individuals to make discriminations finer than 1%. Then if the actual probability is ½%, spending more on enforcement to make the probability 1% would not be beneficial because deterrence would not increase.

[74] Bebchuk and Kaplow (1992) consider imperfect information about the probability of sanctions and emphasize that maximal sanctions may not be socially desirable. See also Kaplow (1990a), which takes into account learning about whether acts are subject to sanctions, and Sah (1991), which focuses on the process by which individuals form perceptions of the probability of detection.

[75] The discussion in this section is based principally on Polinsky and Shavell (2001). We do not examine the corruption of the government procurement process, such as the payment of a bribe to a government agent in order to obtain a defense contract.

a bribe offered by an individual and an extortion demand made by the enforcer—a payment for not turning in the individual.) A second form of corruption is framing and framing-related extortion, in which an enforcement agent may frame an innocent individual or threaten to frame him in order to extort money from him.

One reason bribery is socially undesirable is that it dilutes deterrence of violations of law. This is because bribery results in a lower payment by an individual than the sanction for the offense. To be concrete, let

$\lambda =$ fraction obtained by the enforcer of the surplus from a bribe agreement.

Therefore, because the surplus from a bribe agreement is the fine f, an individual pays a bribe of λf. To the degree that λ is less than one, there is underdeterrence.[76] Similarly, bribery leads to excessive participation in the harm-creating activity.

Framing and framing-related extortion also dilute deterrence of violations of law. The reason is that framing and extortion imply that those who act innocently face an expected sanction, so that the difference between the expected sanction if individuals commit a violation and if they do not is lessened. If, for example, individuals who violate the law face an expected fine of $1,000 and innocent individuals face an expected fine of $200 due to the risk of being extorted or framed, the additional cost to an innocent individual of committing the offense is $800 instead of $1,000. (This point is essentially the same as the observation in section 15 that mistaken convictions dilute deterrence.) Additionally, framing and framing-related extortion undesirably discourage participation in the harm-creating activity.

Because corruption dilutes deterrence and distorts activity decisions, its control may be socially desirable. One way to reduce corruption is to impose fines (or imprisonment sentences) on individuals caught engaging in bribery, extortion, and framing. For example, suppose there is a fine for bribery imposed on the enforcer with a probability. Let

$f_B =$ fine imposed on the enforcer for engaging in bribery; and

$p_B =$ probability that an enforcer is caught engaging in bribery.

Bribery will be deterred if the surplus from a bribe agreement is eliminated, that is, if $p_B f_B \geq f$. Otherwise, the bribe payment will be $p_B f_B + \lambda(f - p_B f_B)$, which exceeds λf, so that deterrence of the offense is greater due to the sanctioning of bribery. For previously discussed reasons, the optimal fine to impose on risk-neutral parties for engaging in bribery is maximal, and the optimal fine for enforcers who frame innocent individuals also is maximal. But, surprisingly, framing-related extortion should not be penalized.[77]

[76] Garoupa and Klerman (2004) discuss how the threat of an imprisonment sanction for the offense will lead the offender to pay a higher bribe than otherwise, thereby reducing the deterrence-diluting effect of bribery.

[77] The kernel of the reason is that penalizing framing-related extortion will lead to one of two detrimental consequences: it will either fail to deter extortion and result in higher costs to innocent individuals (the sum

Corruption also can be reduced by paying enforcers rewards for reporting violations. Such payments will reduce their incentive to accept bribes because they will sacrifice their rewards if they fail to report violations. Indeed, sufficiently high rewards would eliminate all incentives to accept bribes. But high rewards may not be optimal because they give enforcers a greater incentive to frame innocent individuals, and high rewards tend to increase framing-related extortion payments (because enforcers sacrifice more by accepting the extortion payment). The optimal reward balances the beneficial effect of using rewards to offset the dilution of deterrence due to bribery with the detrimental effects associated with increased framing and extortion of innocent individuals.

A third way to control corruption is to pay enforcers more than their reservation wage (that is, to pay them an efficiency wage). Then they would have more to lose if punished for corrupt behavior and denied future employment. There is, however, a social cost to the state of paying enforcers more than the wage necessary to attract them—the distortions caused by the additional taxes needed to make such payments.

The discussion to this point implicitly presumed that the fine for the harmful act is fixed. A question that naturally arises, however, is whether the deterrence-diluting effects of corruption can be offset by raising the fine on offenders. For example, suppose that the optimal fine would be $100 if a fine were always paid when an offender is caught, but that bribery results in a bribe payment equal to $50, one half of the fine. Could not the fine on an offender be increased to $200, so that the bribe would then be $100 and the effective penalty be exactly what is desired? In the basic risk-neutral model of enforcement, it is not possible to raise the fine because the optimal fine is maximal. More realistically, however, the optimal fine is less than maximal for a variety of reasons, including those related to risk aversion, marginal deterrence, and general enforcement. Then, while it would be possible to raise the fine to offset the deterrence-diluting effects of corruption, doing so would lead to social costs (for example, greater bearing of risk) and might not be desirable.[78]

of their expected extortion payment and the expected fine on them for paying extortion); or else it will cause enforcers to switch from extorting money from innocent individuals to actually framing them, which is socially worse. This result is demonstrated in Polinsky and Shavell (2001), which also discusses qualifications to this point.

[78] Becker and Stigler (1974) focus on the control of bribery and consider paying rewards to enforcers or requiring them to post bonds that would be forfeited if they are caught engaging in bribery. Mookherjee and Png (1995) analyze bribery and conclude, given their assumption that fines are unbounded, that it is optimal to eliminate bribery. Bowles and Garoupa (1997) also discuss the control of bribery through sanctions. Hindriks, Keen, and Muthoo (1999) study bribery and extortion in the context of tax evasion, and examine rewards and penalties as methods of control. Other writing on corruption includes Pashigian (1975), Klitgaard (1988), Shleifer and Vishny (1993), Bardhan (1997), and Rose-Ackerman (1999); several of these articles focus on corruption in the awarding of government contracts and licenses rather than corruption in the imposition of sanctions for violations of law.

25. Incapacitation

Our discussion of public enforcement has presumed that the threat of sanctions reduces harm by discouraging individuals from causing harm—that is, by deterring them. However, a different way for society to reduce harm is by imposing sanctions that remove parties from positions in which they are able to cause harm—that is, by *incapacitating* them. Imprisonment is the primary incapacitative sanction, although there are other examples: individuals can lose their drivers licenses, preventing them from doing harm while driving; businesses can lose their right to operate in certain domains, and the like. We focus here on imprisonment, but what we say applies to incapacitative sanctions generally.

To better understand public enforcement when sanctions are incapacitative, suppose that their sole function is to incapacitate; that is, assume for simplicity that sanctions do not deter. (For instance, deterrence might not occur if, given the relevant range of the probabilities and magnitudes of the sanctions, individuals' gains from harmful acts exceed the expected sanctions.) Let

$$h(t) = \text{harm that would be caused by an individual at age } t \text{ if not in jail.}$$

We assume that $h(t)$ either is constant or declines with age.

Assuming that the social goal is to minimize the sum of the harm and the cost c of incarceration,[79] the optimal policy is to keep an individual in jail as long as $h(t) > c$. In other words, if the harm the individual would cause exceeds the cost of imprisonment, an individual should be put in prison and kept there as long as the harm continues to exceed the cost of imprisonment. He should be released otherwise. Put differently, the optimal prison term as a function of potential harm caused is zero up to a threshold—the point at which harm equals the cost of imprisonment—and then rises discontinuously to the length of time during which the person's harm if released would exceed imprisonment costs. Jail should only be used to incapacitate individuals who otherwise would have caused relatively high harm.

Two points about the incapacitative rationale are important to note. First, evidence exists suggesting that the harm caused by individuals declines with their age.[80] Thus, from the incapacitative standpoint, it often will be desirable to release older prisoners from jail. Second, as a matter of logic, the incapacitative rationale might imply that a person should be put in jail even if he has not committed a crime—if his danger to society makes incapacitating him worthwhile. This would be true, for example, if there were some accurate way to predict a person's dangerousness independently of his actual behavior. In practice, however, the fact that a person has committed a harmful act may be

[79] For simplicity, we are not taking into account here the gains that individuals obtain from committing offenses or the disutility that individuals bear from time in jail.

[80] See, for example, Wilson and Herrnstein (1985, pp. 126–147) and U.S. Department of Justice (1997a, pp. 371, table 4.4, 378–379, table 4.7).

a good basis for predicting his future behavior, in which case the incapacitation rationale would imply that a jail term should be imposed only if the individual has committed an especially harmful act.

The optimal probability of detection is determined by a straightforward tradeoff. The higher the probability, the greater the number of individuals who will be incapacitated, resulting in social gains equal to the difference between the harm that individuals would cause and the cost of their incapacitation. But the higher the probability, the higher are enforcement costs. At some point, it is optimal to stop raising the probability, when the marginal social gains just equal the marginal cost of raising the probability.

Last, we briefly comment on the relationship between the nature of optimal enforcement when incapacitation is the goal versus when deterrence is the goal. First, when incapacitation is the goal, the optimal length of the prison term (which is determined by the condition that $h(t) > c$) is independent of the probability of apprehension. In contrast, when deterrence is the goal, the optimal sanction depends on the probability—the sanction generally is higher the lower is the probability. Second, when incapacitation is the goal, the probability and magnitude of sanctions are independent of the ability to deter. Thus, for example, if this ability is limited (consider individuals who commit crimes while enraged), a low expected sanction may be optimal under the deterrence rationale, but a high expected sanction still might be called for to incapacitate.[81]

26. Costly observation of wealth

In our prior discussions of optimal sanctions, we implicitly assumed that the enforcement authority could costlessly observe individuals' wealth levels. Knowing wealth levels, the enforcement authority then chose the sanctions to impose, fines and/or imprisonment sentences. In fact, however, individuals and firms may be able to hide assets from government enforcers, including by hoarding cash, transferring assets to relatives or related legal entities, or moving money to offshore bank accounts. In this section we consider optimal sanctions when an individual's level of wealth can be observed only after a costly audit or not at all.[82]

Suppose first that the enforcement authority employs fines as sanctions and can audit an individual who claims that he cannot pay the fine. If the audit determines that the

[81] See Shavell (1987a) for a theoretical examination of optimal incapacitation policy, Ehrlich (1981, pp. 315–316, 319–321) for a model used to estimate the relative importance of incapacitation and deterrence, and Levitt (1998) and Kessler and Levitt (1999) for empirical studies of incapacitation and deterrence. Economists have paid much less attention to incapacitation than to deterrence, despite the significance of the incapacitation rationale in criminal law enforcement.

[82] The discussion in this section is based on Polinsky (2006a, 2006b). See also Chu and Jiang (1993) and Levitt (1997), who consider the choice between fines and imprisonment when wealth cannot be discovered by the enforcement authority at any cost (in Chu and Jiang's case, this assumption is implicit), and Garoupa (1998), who investigates optimal fines when the enforcement authority is assumed to costlessly observe an underestimate of offenders' wealth levels.

individual misrepresented his wealth level, he can be fined for having lied about his wealth. Assume for simplicity that this fine is independent of the degree of misrepresentation. The enforcement authority's problem is to choose, so as to maximize social welfare, the probability of detecting the offense, the fine for the offense, the probability of an audit conditional on the individual's claiming that he cannot pay the fine, and the fine for misrepresentation of wealth. Without loss of generality, we assume that if the latter fine is applicable, it is imposed instead of the fine for the offense (rather than in addition to the fine for the offense).

It can be demonstrated that the optimal fine for misrepresenting one's wealth level equals the fine for the offense divided by the audit probability, and therefore generally exceeds the fine for the offense. This is a natural generalization of the formula for the optimal fine, given the probability of detection, which is the harm divided by the probability. In effect, the "harm" from the act of concealing one's wealth is the failure to pay the fine for the original offense, so the optimal fine for concealment is this harm divided by the applicable probability of being caught if one engages in it.

Assuming the harmful act is worth controlling, the optimal audit probability is positive, increases as the cost of an audit declines, and equals one if the cost is sufficiently low. Auditing is valuable because it reduces misrepresentation of wealth and thereby increases deterrence. If the optimal audit probability is less than one, however, some individuals who are capable of paying the fine for the offense will misrepresent their wealth levels. Unlike in the basic analysis in which wealth is assumed to be costlessly observable, the optimal fine for the offense now results in underdeterrence, due to the cost of auditing wealth levels. By reducing the fine for the offense from the level that would lead to first-best behavior, fewer individuals will misrepresent their wealth levels, so auditing costs decline. The reduction in deterrence has no first-order effect on social welfare because the marginal individuals who are induced to commit the offense have gains equal to the harm from the offense.

Next suppose that the enforcement authority simply cannot observe wealth, say because the cost of an audit is prohibitively high.[83] If the enforcement authority would have used fines alone if it could have observed wealth at no cost, it would have imposed a higher fine on higher-wealth individuals.[84] It obviously cannot do this now. Instead, it may be desirable to use the threat of an imprisonment sentence to induce individuals capable of paying a higher fine to do so. Specifically, if there are two levels of wealth, the enforcement authority might set the fine greater than the wealth level of the low-wealth individuals and impose an imprisonment sentence on any individuals who do not pay this fine. By using an imprisonment sentence in this way, high-wealth individuals can be induced to pay a higher fine, which is socially beneficial, though low-wealth individuals now will incur a socially costly sanction, imprisonment. In other

[83] The following discussion introduces imprisonment as an alternative, or supplement, to fines. For simplicity, imprisonment was not considered above in the discussion of auditing.

[84] For now familiar reasons, it would have done this to reduce or eliminate the underdeterrence that otherwise would occur if the fine were the same as for low-wealth individuals.

words, when wealth is not observable, it may be desirable to impose a costly sanction—imprisonment sentences—on low-wealth individuals in order to better deter high-wealth offenders through a cheap sanction—fines.

Another possibility is that the enforcement authority would have used both fines and imprisonment if it could have observed wealth at no cost. Perhaps surprisingly, the inability to observe wealth is of no consequence in this case. The reason, in essence, is that the mix of fines and imprisonment that would be chosen when wealth is observable will impose a higher burden (though a lower fine) on low-wealth individuals. Thus, high-wealth individuals will naturally want to identify themselves. Specifically, they will prefer to pay a higher fine and bear a shorter imprisonment sentence than to masquerade as low-wealth individuals, who will bear longer imprisonment sentences and a higher overall burden. Consequently, the same mix of sanctions that would have been imposed on both groups if wealth were costlessly observable can be used when wealth is not observable.

In summary, information about wealth levels is useful only if the enforcement authority would want to impose a higher burden of sanctions on high-wealth individuals than on low-wealth individuals, for then high-wealth individuals would pretend to be low-wealth individuals if wealth could not be observed. This is the case if the enforcement authority would want to use fines alone. If, however, the enforcement authority would want to impose a lower burden of sanctions on high-wealth individuals than on low-wealth individuals, high-wealth individuals will voluntarily bear such sanctions even if they include a higher fine. This case is applicable when the enforcement authority would want to use fines and imprisonment together. Monitoring of wealth levels may be worthwhile in the first case, but is not needed in the second case.

27. Social norms

Although we have restricted attention to public enforcement of law, social norms and morality should be mentioned because they influence in significant ways the attainment of desired behavior.[85] By social norms (or moral rules), we mean conduct—such as keeping promises, not lying, and not harming others—that is associated with certain distinctive psychological and social attributes. In particular, social norms influence behavior partly through internal incentives: when a person obeys a moral rule, he will tend to feel virtuous, and if he disobeys the rule, he will tend to feel guilty. Social norms also influence behavior through external incentives: when a person is observed by another party to have obeyed a moral rule, that party may bestow praise on the first party, who will enjoy the praise; and if the person is observed by the other party to have disobeyed the rule, the second party will tend to disapprove of the first party, who will dislike the disapproval.

[85] On social norms and the law, see generally McAdams and Rasmusen (2007). See also University of Pennsylvania Law Review (1996) and Posner (1997).

Because social norms affect behavior through the foregoing moral incentives, some socially desirable conduct can be encouraged reasonably well without employing the legal system.[86] For example, whether an individual cuts in line at the movie theater, keeps his lunch engagements, or lets his children make a nuisance of themselves at the supermarket, is controlled with rough success by internal and external moral incentives. Such conduct generally can be regulated satisfactorily by moral incentives alone because, among other things, internal incentives work automatically (we know when we have done the wrong thing), external incentives are likely to apply (if a person cuts in line, this will be noticed), and the benefits from violations are not large, so that they can be outweighed relatively easily by the force of the moral incentives.

The need for formal law enforcement stems from two principal considerations. First, much conduct that society desires cannot be controlled through moral incentives alone. One reason is that the private gains from undesirable conduct are often large. The utility obtained by a robber, a tax cheat, or a polluter may be substantial, and dominate the moral incentives. Another reason is that external moral sanctions might be imposed only with a low probability (the robber, tax cheat, or polluter might not be spotted by others). A second rationale for formal law enforcement is that the social harm from failing to control an act through moral incentives may be large. This makes the expense of law enforcement worth incurring (as in the case of robbery, but not of cutting in line at movie theaters).

Although we have been treating social norms and formal law enforcement as distinct ways of controlling behavior, law enforcement might also influence social norms.[87] For instance, enforcement of laws prohibiting discrimination based on race may change beliefs about proper conduct (reinforcing internal moral incentives) and lead to a greater willingness of individuals to express disapproval when they witness discriminatory behavior (enhancing external moral incentives). The importance of the effect of law on social norms may be limited, however, to the extent that social norms are mainly the result of early childhood experience and the messages conveyed by parents and other authority figures, such as educators.

28. Fairness

To this point we have not considered the possibility that individuals have opinions about the fairness of sanctions or the arbitrariness of enforcement.[88]

[86] For a comparison of social norms and law enforcement as means of controlling behavior, see Shavell (2002).

[87] See, for example, McAdams (1997) and Sunstein (1996).

[88] The discussion in this section is based on Polinsky and Shavell (2000b) and Kaplow and Shavell (2002, pp. 291–378), the latter of which relates the economic analysis of fairness in enforcement to the philosophical literature. See also Miceli (1991) and Diamond (2002), who derive optimal sanctions taking both deterrence and the fairness of sanctions into account, but holding the probability of sanctions fixed. Waldfogel (1993) studies empirically whether actual sanctions are better explained by considerations of deterrence or fairness.

Suppose, first, that individuals believe that sanctions should be imposed on those who have committed certain bad acts and that the magnitude of sanctions should reflect the gravity of the acts. A formal assumption that captures this view is that individuals obtain fairness-related utility from the imposition of sanctions on those who committed the bad acts, where this utility is maximized at a *fairness-ideal* level of sanction that depends on the harmfulness (or a related aspect) of the acts. The fairness-ideal sanction may be lower or higher than the conventionally optimal sanction that we have discussed above. Note in particular that the fairness-ideal sanction does not depend on the probability of detection, whereas the conventionally optimal sanction is higher the lower is the probability of detection, suggesting that if the probability of detection is sufficiently low, the conventionally optimal sanction will exceed the fairness-ideal sanction. In any case, when fairness-related utility is taken into account along with the other elements of social welfare considered above, the optimal sanction will be a compromise between the fairness-ideal sanction and the conventionally optimal sanction.

When both the probability and magnitude of sanctions may be varied, the conventional solution to the enforcement problem also is altered because of fairness considerations. As discussed previously, if individuals are risk neutral, the usual solution consists of the highest possible sanction and a relatively low probability. When the issue of fairness is added to the analysis, however, the usual solution generally is not optimal because a very high sanction will be seen as unfair, or more precisely, will result in the lowering of individuals' fairness-related utility. With respect to double parking, for example, even a sanction of $100 might be considered unfair because double parking is regarded as only a slightly bad act.

A consequence of the desire to keep sanctions at fair levels, meaning quite constrained levels for acts that are not very bad or harmful, is that the socially optimal probability of detection changes. One possibility is that the optimal probability would be higher, perhaps much higher than the conventionally optimal probability: to achieve a desired level of deterrence with a lower fairness-restricted sanction, the probability has to rise. If the sanction for double parking cannot exceed $100 because of fairness considerations, then, to create an expected sanction of $10, the probability must be 10%, greatly exceeding the approximately $1/10\%$ probability that would be optimal if risk-neutral individuals have wealth of $10,000 and the fine is set at this level. Another possibility, though, is that the optimal probability would be lower than in the conventional case: the additional deterrence created by raising the probability might be relatively low because the sanction is relatively low; and the lower the deterrent benefit from raising the probability, the lower would be the social incentive to devote resources to enforcement.

Another concept of fairness concerns the probability of detection rather than the magnitude of sanctions. Suppose that individuals consider it unfair for some violators of law to be sanctioned when others who were lucky enough not to be caught are not sanctioned. Specifically, suppose that individuals experience fairness-related disutility if there is only a probability rather than a certainty of sanctions, and their disutility rises the lower the probability of detection. Then the optimal probability would be higher, and therefore the optimal sanction would be lower, than in the absence of this fairness-related component of social welfare.

A further notion of fairness involves the form of liability, whether liability is strict or based on fault. Individuals might prefer fault-based liability because sanctions are imposed on parties only if they behaved in a socially inappropriate way. If individuals derive greater fairness-related utility from use of fault-based liability, then this form of liability is more likely to be optimal than we have suggested previously.

A final issue concerns the relevance of fairness considerations when firms, as opposed to individuals, are sanctioned. If sanctions are imposed on firms, then fairness-related utility may have to be reconsidered, presuming that what matters in terms of fairness is that the *individuals* responsible for harmful acts bear sanctions as opposed to the artificial legal entity of a firm. Specifically, one would want to identify the sanctions actually suffered by such persons within a firm if the firm bears a sanction. For example, if a firm demotes a person who negligently caused the firm to incur a sanction, then the person's loss from the demotion would be the fairness-relevant sanction, not the sanction imposed on the firm. Note, too, that the imposition of sanctions on firms often penalizes individuals who are unlikely to be considered responsible for the harm, namely shareholders and customers. To the extent that the fairness goal is to penalize responsible individuals within firms, and not firms as entities, the social value of sanctions in achieving fairness is attenuated.

29. Conclusion

Having reviewed the theory of public enforcement of law, we want to conclude by commenting on two types of private behavior that bear significantly on public law enforcement.

First, private parties may themselves take actions to prevent being harmed. For example, to reduce the risk of being criminally victimized, individuals might install locks on their possessions, carry weapons, or hire security personnel. These private efforts to prevent or deter harmful acts serve as a partial substitute for public efforts; moreover, private efforts are sometimes more efficient than public efforts (citizens may know better where to put locks), though they also may be less efficient (public authorities may know better how to assign police). These observations raise important questions about the relationship between private protection and public enforcement. Should private efforts to protect against being victimized be regulated? Does the state spend too little on public enforcement, relying on the fact that private actors often undertake their own defensive efforts? The optimal relationship between private and public efforts to control harmful activities deserves more careful examination.[89]

[89] There is some literature that discusses the issues raised in this paragraph; see, for example, Clotfelter (1977, 1978) and Shavell (1991a).

Second, private individuals may bring suits against parties who also may be subject to publicly imposed penalties. For example, a victim of an automobile accident may sue the driver who caused the accident, and this driver might also be sanctioned by the state for a driving infraction. Private lawsuits channel harm-creating behavior and thus constitute a substitute, at least to some extent, for public enforcement. This leads one to ask how public enforcement and parallel private litigation should be managed. Should the payment of a public penalty be an offset to private damages, and vice versa? Should the state regulate private litigation so as to better coordinate it with public enforcement? Is public enforcement or private litigation the socially cheaper way to accomplish desired behavior?

A full treatment of the control of harm-creating behavior would address both sets of issues just discussed.

References

Andreoni, J. (1991). "Reasonable Doubt and the Optimal Magnitude of Fines: Should the Penalty Fit the Crime?" RAND Journal of Economics 22, 385–395.

Andreoni, J., Erard, B., Feinstein, J. (1998). "Tax Compliance". Journal of Economic Literature 36, 818–860.

Arlen, J.A. (1994). "The Potentially Perverse Effects of Corporate Criminal Liability". Journal of Legal Studies 23, 833–867.

Bardhan, P. (1997). "Corruption and Development: A Review of Issues". Journal of Economic Literature 35, 1320–1346.

Bar-Gill, O., Harel, A. (2001). "Crime Rates and Expected Sanctions: The Economics of Deterrence Revisited". Journal of Legal Studies 30, 485–501.

Bebchuk, L.A., Kaplow, L. (1992). "Optimal Sanctions When Individuals Are Imperfectly Informed About the Probability of Apprehension". Journal of Legal Studies 21, 365–370.

Bebchuk, L.A., Kaplow, L. (1993). "Optimal Sanctions and Differences in Individuals' Likelihood of Avoiding Detection". International Review of Law and Economics 13, 217–224.

Beccaria, C. (1767). On Crimes and Punishments, and Other Writings. Cambridge University Press, New York. Editor, R. Bellamy, Translator, R. Davies et al. 1995.

Becker, G.S. (1968). "Crime and Punishment: An Economic Approach". Journal of Political Economy 76, 169–217.

Becker, G.S., Stigler, G.J. (1974). "Law Enforcement, Malfeasance, and Compensation of Enforcers". Journal of Legal Studies 3, 1–18.

Bentham, J. (1789). An Introduction to the Principles of Morals and Legislation. In: The Utilitarians. Anchor Books, Garden City, N.Y., 1973.

Block, M.K., Sidak, J.G. (1980). "The Cost of Antitrust Deterrence: Why Not Hang a Price Fixer Now and Then?" Georgetown Law Journal 68, 1131–1139.

Bouckaert, B., De Geest, G. (Eds.) (1992). Bibliography of Law and Economics. Kluwer Academic Publishers, Dordrecht.

Bowles, R., Garoupa, N. (1997). "Casual Police Corruption and the Economics of Crime". International Review of Law and Economics 17, 75–87.

Carr-Hill, R.A., Stern, N.H. (1979). Crime, the Police and Criminal Statistics. Academic Press, London.

Chu, C.Y.C., Hu, S., Huang, T. (2000). "Punishing Repeat Offenders More Severely". International Review of Law and Economics 20, 127–140.

Chu, C.Y.C., Jiang, N. (1993). "Are Fines More Efficient than Imprisonment?" Journal of Public Economics 51, 391–413.

Clotfelter, C.T. (1977). “Public Services, Private Substitutes, and the Demand for Protection Against Crime”. American Economic Review 67, 867–877.

Clotfelter, C.T. (1978). “Private Security and the Public Safety”. Journal of Urban Economics 5, 388–402.

Cohen, M.A. (1989). “Corporate Crime and Punishment: A Study of Social Harm and Sentencing Practice in the Federal Courts, 1984–1987”. American Criminal Law Review 26, 605–660.

Cohen, M.A. (1999). “Monitoring and Enforcement of Environmental Policy”. In: Folmer, H., Tietenberg, T. (Eds.), The International Yearbook of Environmental and Resource Economics 1999/2000. Edward Elgar, Cheltenham, United Kingdom, pp. 44–106.

Cooter, R.D., Rubinfeld, D.L. (1989). “Economic Analysis of Legal Disputes and Their Resolution”. Journal of Economic Literature 27, 1067–1097.

Craswell, R., Calfee, J.E. (1986). “Deterrence and Uncertain Legal Standards”. Journal of Law, Economics, & Organization 2, 279–303.

Diamond, P.A. (2002). “Integrating Punishment and Efficiency Concerns in Punitive Damages for Reckless Disregard of Risks to Others”. Journal of Law, Economics, & Organization 18, 117–139.

Eide, E. (2000). “Economics of Criminal Behavior”. In: Bouckaert, B., De Geest, G. (Eds.), Encyclopedia of Law and Economics. Edward Elgar Publishing Limited, Cheltenham, United Kingdom, pp. 345–389.

Ehrlich, I. (1981). “On the Usefulness of Controlling Individuals: An Economic Analysis of Rehabilitation, Incapacitation and Deterrence”. American Economic Review 71, 307–332.

Emons, W. (2003). “A Note on the Optimal Punishment for Repeat Offenders”. International Review of Law and Economics 23, 253–259.

Friedman, D.D. (1995). “Making Sense of English Law Enforcement in the Eighteenth Century”. University of Chicago Law School Roundtable 2, 475–505.

Garoupa, N. (1997). “The Theory of Optimal Law Enforcement”. Journal of Economic Surveys 11, 267–295.

Garoupa, N. (1998). “Optimal Law Enforcement and Imperfect Information when Wealth Varies Among Individuals”. Economica 65, 479–490.

Garoupa, N., Klerman, D. (2004). “Corruption and the Optimal Use of Nonmonetary Sanctions”. International Review of Law and Economics 24, 219–225.

Glaeser, E.L., Kessler, D.P., Piehl, A.M. (2000). “What Do Prosecutors Maximize? An Analysis of the Federalization of Drug Crimes”. American Law and Economics Review 2, 259–290.

Grossman, G.M., Katz, M.L. (1983). “Plea Bargaining and Social Welfare”. American Economic Review 73, 749–757.

Hindriks, J., Keen, M., Muthoo, A. (1999). “Corruption, Extortion and Evasion”. Journal of Public Economics 74, 395–430.

Innes, R. (1999). “Remediation and Self-Reporting in Optimal Law Enforcement”. Journal of Public Economics 72, 379–393.

Kahneman, D., Slovic, P., Tversky, A. (Eds.) (1982). Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge and New York.

Kaplow, L. (1990a). “Optimal Deterrence, Uninformed Individuals, and Acquiring Information About Whether Acts are Subject to Sanctions”. Journal of Law, Economics, & Organization 6, 93–128.

Kaplow, L. (1990b). “A Note on the Optimal Use of Nonmonetary Sanctions”. Journal of Public Economics 42, 245–247.

Kaplow, L. (1992). “The Optimal Probability and Magnitude of Fines for Acts That Definitely Are Undesirable”. International Review of Law and Economics 12, 3–11.

Kaplow, L., Shavell, S. (1994a). “Accuracy in the Determination of Liability”. Journal of Law and Economics 37, 1–15.

Kaplow, L., Shavell, S. (1994b). “Optimal Law Enforcement with Self-Reporting of Behavior”. Journal of Political Economy 102, 583–606.

Kaplow, L., Shavell, S. (2002). Fairness Versus Welfare. Harvard University Press, Cambridge.

Kenkel, D.S. (1993). “Do Drunk Drivers Pay Their Way? A Note on Optimal Penalties for Drunk Driving”. Journal of Health Economics 12, 137–149.

Kessler, D., Levitt, S.D. (1999). “Using Sentence Enhancements to Distinguish between Deterrence and Incapacitation”. Journal of Law and Economics 42, 343–363.

Klitgaard, R. (1988). Controlling Corruption. University of California Press, Berkeley.

Kobayashi, B.H., Lott, J.R. Jr. (1992). "Low-Probability-High-Penalty Enforcement Strategies and the Efficient Operation of the Plea-Bargaining System". International Review of Law and Economics 12, 69–77.

Kornhauser, L.A. (1982). "An Economic Analysis of the Choice Between Enterprise and Personal Liability for Accidents". California Law Review 70, 1345–1392.

Laffont, J., Tirole, J. (1993). The Theory of Incentives in Procurement and Regulation. MIT Press, Cambridge, MA.

Landes, W.M. (1971). "An Economic Analysis of the Courts". Journal of Law and Economics 4, 61–107.

Landes, W.M., Posner, R.A. (1975). "The Private Enforcement of Law". Journal of Legal Studies 4, 1–46.

Lando, H., Shavell, S. (2004). "The Advantage of Focusing Law Enforcement Effort". International Review of Law and Economics 24, 209–218.

Levitt, S.D. (1997). "Incentive Compatibility Constraints as an Explanation for the Use of Prison Sentences Instead of Fines". International Review of Law and Economics 17, 179–192.

Levitt, S.D. (1998). "Why Do Increased Arrest Rates Appear to Reduce Crime: Deterrence, Incapacitation, or Measurement Error?" Economic Inquiry 36, 353–372.

Levitt, S.D., Miles, T. (2007). "Empirical Study of Criminal Punishment". In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics, vol. 1. North-Holland, Amsterdam.

Lewin, J.L., Trumbull, W.N. (1990). "The Social Value of Crime?" International Review of Law and Economics 10, 271–284.

McAdams, R.H. (1997). "The Origin, Development, and Regulation of Norms". Michigan Law Review 96, 338–433.

McAdams, R.H., Rasmusen, E. (2007). "Norms in Law and Economics". In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics, vol. 2. North-Holland, Amsterdam.

Malik, A.S. (1990). "Avoidance, Screening and Optimum Enforcement". RAND Journal of Economics 21, 341–353.

Malik, A.S. (1993). "Self-Reporting and the Design of Policies for Regulating Stochastic Pollution". Journal of Environmental Economics and Management 24, 241–257.

Miceli, T.J. (1991). "Optimal Criminal Procedure: Fairness and Deterrence". International Review of Law and Economics 11, 3–10.

Miceli, T.J. (1996). "Plea Bargaining and Deterrence: An Institutional Approach". European Journal of Law and Economics 3, 249–264.

Miceli, T.J., Bucci, C. (2005). "A Simple Theory of Increasing Penalties for Repeat Offenders". Review of Law & Economics 1, 71–80. (Available at http://www.bepress.com/rle/vol1/iss1/art5.)

Montesquieu, C.-L. (1748). The Spirit of the Laws. University of California Press, Berkeley, Rept. edn. 1977.

Mookherjee, D. (1997). "The Economics of Enforcement". In: Bose, A., Rakshit, M., Sinha, A. (Eds.), Issues in Economic Theory and Policy, Essays in Honor of Tapas Majumdar. Oxford University Press, New Delhi, pp. 202–249.

Mookherjee, D., Png, I.P.L. (1992). "Monitoring vis-à-vis Investigation in Enforcement of Law". American Economic Review 82, 556–565.

Mookherjee, D., Png, I.P.L. (1994). "Marginal Deterrence in Enforcement of Law". Journal of Political Economy 102, 1039–1066.

Mookherjee, D., Png, I.P.L. (1995). "Corruptible Law Enforcers: How Should They Be Compensated?" Economic Journal 105, 145–159.

Newman, H.A., Wright, D.W. (1990). "Strict Liability in a Principal-Agent Model". International Review of Law and Economics 10, 219–231.

Pashigian, B.P. (1975). "On the Control of Crime and Bribery". Journal of Legal Studies 4, 311–326.

Png, I.P.L. (1986). "Optimal Subsidies and Damages in the Presence of Judicial Error". International Review of Law and Economics 6, 101–105.

Polinsky, A.M. (1980a). "Private versus Public Enforcement of Fines". Journal of Legal Studies 9, 105–127.

Polinsky, A.M. (1980b). "Strict Liability vs. Negligence in a Market Setting". American Economic Review: Papers and Proceedings 70, 363–370.

Polinsky, A.M. (2006a). "The Optimal Use of Fines and Imprisonment When Wealth is Unobservable". Journal of Public Economics 90, 823–835.

Polinsky, A.M. (2006b). "Optimal Fines and Auditing When Wealth is Costly to Observe". International Review of Law and Economics 26, 323–335.

Polinsky, A.M., Rubinfeld, D.L. (1988). "The Deterrent Effects of Settlements and Trials". International Review of Law and Economics 8, 109–116.

Polinsky, A.M., Rubinfeld, D.L. (1989). "A Note on Optimal Public Enforcement with Settlements and Litigation Costs". Research in Law and Economics 12, 1–8.

Polinsky, A.M., Rubinfeld, D.L. (1991). "A Model of Optimal Fines for Repeat Offenders". Journal of Public Economics 46, 291–306.

Polinsky, A.M., Shavell, S. (1979). "The Optimal Tradeoff between the Probability and Magnitude of Fines". American Economic Review 69, 880–891.

Polinsky, A.M., Shavell, S. (1984). "The Optimal Use of Fines and Imprisonment". Journal of Public Economics 24, 89–99.

Polinsky, A.M., Shavell, S. (1991). "A Note on Optimal Fines When Wealth Varies Among Individuals". American Economic Review 81, 618–621.

Polinsky, A.M., Shavell, S. (1992). "Enforcement Costs and the Optimal Magnitude and Probability of Fines". Journal of Law and Economics 35, 133–148.

Polinsky, A.M., Shavell, S. (1993). "Should Employees Be Subject to Fines and Imprisonment Given the Existence of Corporate Liability?" International Review of Law and Economics 13, 239–257.

Polinsky, A.M., Shavell, S. (1998). "On Offense History and the Theory of Deterrence". International Review of Law and Economics 18, 305–324.

Polinsky, A.M., Shavell, S. (1999). "On the Disutility and Discounting of Imprisonment and the Theory of Deterrence". Journal of Legal Studies 28, 1–16.

Polinsky, A.M., Shavell, S. (2000a). "The Economic Theory of Public Enforcement of Law". Journal of Economic Literature 38, 45–76.

Polinsky, A.M., Shavell, S. (2000b). "The Fairness of Sanctions: Some Implications for Optimal Enforcement Policy". American Law and Economics Review 2, 223–237.

Polinsky, A.M., Shavell, S. (2001). "Corruption and Optimal Law Enforcement". Journal of Public Economics 81, 1–24.

Posner, R.A. (1985). "An Economic Theory of the Criminal Law". Columbia Law Review 85, 1193–1231.

Posner, R.A. (1997). "Social Norms and the Law: An Economic Approach". American Economic Review: Papers and Proceedings 87, 365–369.

Reinganum, J.F. (1988). "Plea Bargaining and Prosecutorial Discretion". American Economic Review 78, 713–728.

Rose-Ackerman, S. (1999). Corruption and Government: Causes, Consequences, and Reform. Cambridge University Press, Cambridge.

Rubinstein, A. (1979). "An Optimal Conviction Policy for Offenses that May Have Been Committed by Accident". In: Brams, S.J., Schotter, A., Schwödiauer, G. (Eds.), Applied Game Theory. Physica-Verlag, Wurzburg, pp. 406–413.

Sah, R.K. (1991). "Social Osmosis and Patterns of Crime". Journal of Political Economy 99, 1272–1295.

Segerson, K., Tietenberg, T. (1992). "The Structure of Penalties in Environmental Enforcement: An Economic Analysis". Journal of Environmental Economics and Management 23, 179–200.

Shavell, S. (1980). "Strict Liability versus Negligence". Journal of Legal Studies 9, 1–25.

Shavell, S. (1985). "Criminal Law and the Optimal Use of Nonmonetary Sanctions as a Deterrent". Columbia Law Review 85, 1232–1262.

Shavell, S. (1987a). "A Model of Optimal Incapacitation". American Economic Review: Papers and Proceedings 77, 107–110.

Shavell, S. (1987b). "The Optimal Use of Nonmonetary Sanctions as a Deterrent". American Economic Review 77, 584–592.

Shavell, S. (1991a). "Individual Precautions to Prevent Theft: Private versus Socially Optimal Behavior". International Review of Law and Economics 11, 123–132.

Shavell, S. (1991b). "Specific versus General Enforcement of Law". Journal of Political Economy 99, 1088–1108.

Shavell, S. (1992). "A Note on Marginal Deterrence". International Review of Law and Economics 12, 345–355.

Shavell, S. (1993). "The Optimal Structure of Law Enforcement". Journal of Law and Economics 36, 255–287.

Shavell, S. (1997a). "The Fundamental Divergence Between the Private and the Social Motive to Use the Legal System". Journal of Legal Studies 26, 575–612.

Shavell, S. (1997b). "The Optimal Level of Corporate Liability Given the Limited Ability of Corporations to Penalize Their Employees". International Review of Law and Economics 17, 203–213.

Shavell, S. (2002). "Law versus Morality as Regulators of Conduct". American Law and Economics Review 4, 227–257.

Shavell, S. (2004). Foundations of Economic Analysis of Law. Harvard University Press, Cambridge, MA.

Shleifer, A., Vishny, R.W. (1993). "Corruption". Quarterly Journal of Economics 108, 599–617.

Spier, K.E. (1997). "A Note on the Divergence Between the Private and the Social Motive to Settle Under a Negligence Rule". Journal of Legal Studies 26, 613–621.

Spier, K.E. (2007). "Litigation". In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics, vol. 1. North-Holland, Amsterdam.

Stigler, G.J. (1970). "The Optimum Enforcement of Laws". Journal of Political Economy 78, 526–536.

Sunstein, C.R. (1996). "Social Norms and Social Roles". Columbia Law Review 96, 201–266.

Sykes, A.O. (1981). "An Efficiency Analysis of Vicarious Liability Under the Law of Agency". Yale Law Journal 91, 168–206.

University of Pennsylvania Law Review (1996). "Symposium: Law, Economics, and Norms". University of Pennsylvania Law Review 144, 1643–2339.

U.S. Department of Justice (1988). Report to the Nation on Crime and Justice, Technical Appendix, 2nd edn. U.S. Department of Justice, Bureau of Justice Statistics, Washington, D.C., NCJ-112011.

U.S. Department of Justice (1997a). Sourcebook of Criminal Justice Statistics—1996. U.S. Department of Justice, Office of Justice Statistics, Bureau of Justice Statistics, Washington, D.C., NCJ-165361.

U.S. Department of Justice (1997b). Uniform Crime Reports for the United States, 1996. U.S. Department of Justice, Federal Bureau of Investigation, Washington, D.C.

U.S. Department of Justice (1998). Profile of Jail Inmates 1996. U.S. Department of Justice, Office of Justice Programs, Washington D.C., NCJ-164620.

Waldfogel, J. (1993). "Criminal Sentences as Endogenous Taxes: Are They 'Just' or 'Efficient'?" Journal of Law and Economics 36, 139–151.

Wilde, L.L. (1992). "Criminal Choice, Nonmonetary Sanctions, and Marginal Deterrence: A Normative Analysis". International Review of Law and Economics 12, 333–344.

Wilson, J.Q., Herrnstein, R.J. (1985). Crime and Human Nature. Simon and Schuster, New York.

Chapter 7

EMPIRICAL STUDY OF CRIMINAL PUNISHMENT

STEVEN D. LEVITT

Department of Economics, University of Chicago, and American Bar Foundation

THOMAS J. MILES

School of Law, University of Chicago

Contents

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)01007-9

Abstract

This chapter reviews empirical studies of criminal punishment and the criminal justice system by economists. Since the modern exposition of the economic model of criminal behavior, empirical economists have tested its predictions using variation in expected criminal punishments. In the past decade, empirical economists have made substantial progress in identifying the effects of punishment on crime by finding new ways to break the simultaneity of crime rates and punishments. The new empirical evidence generally supports the deterrence model but shows that incapacitation influences crime rates, too. Evidence of the crime-reducing effect of the scale of policing and incarceration is consistent across different methodological approaches. Estimates of the deterrent effect of other penalties, such as capital punishment, are less robust and suggest that claims of large effects from these policies may be spurious. More work is needed to assess the relative importance of deterrence and incapacitation.

Empirical economists have made less progress in studying the criminal justice system itself. Data availability constrains the economists' ability to resolve the simultaneity of criminal justice institutions and crime rates, and this limitation hampers rigorous empirical investigation of the incentive effects of numerous criminal justice institutions and procedures. This area remains an important avenue for future empirical research.

JEL classification: K14, K42

1. Introduction

Many factors influence the decision to commit crime, the types of crime committed, and the amount of criminal activity undertaken. In most theories of criminal behavior, but especially the rational actor/economic model of crime developed by Becker (1968), publicly administered law enforcement plays a critical role. The chapter by Professors Polinsky and Shavell in this volume reviews the large theoretical literature that Becker's influential article engendered. Briefly, in the Beckerian or rational choice model, a potential offender compares the expected costs and benefits of engaging in criminal activity. He engages in criminal activity when the commission of a crime increases his expected utility by more than the expected sanction. In this framework, an expected sanction in excess of the expected gain dissuades the potential offender, and the model is therefore one of deterrence rather than alternative channels through which punishment may affect crime, such as incapacitation.

This chapter is devoted to understanding empirical economic research to-date on the impact and operation of the criminal justice system. The chapter consists of two primary sections. After a brief review of the theoretical models of crime, one section focuses on efforts to test empirically the economic model of criminal behavior. These studies typically compare movements in crime rates to variation in proxies for expected punishments, such as the size of police forces or incarcerated populations. Three themes emerge from these studies. First, empirical tests of the model are hampered by simultaneity problems. When crime rates influence the formation of public policies to combat crime, standard statistical approaches do not capture the causal impact of these policies. Second, even when a causal relationship between a public policy and the incidence of criminal activity is identified, an attendant question is whether the causal effect is attributable to deterrence or incapacitation. Third, despite these difficulties, researchers have enjoyed significant progress in recent years in testing the economic model. They have found that deterrence has a substantial but far from complete role in explaining observed patterns of criminal activity.[1]

The next section of this chapter reviews empirical studies of particular aspects of the criminal justice system itself. These studies typically relate the treatment that a defendant receives from the criminal justice system to the defendant's personal characteristics. To date, relatively few of these studies have attempted to link particular criminal justice practices or procedures to the incidence of criminal activity. Studies in this literature differ from tests of the economic model of crime in two ways. They are

[1] Because the focus of this chapter is on the impact of public enforcement of the law on criminal activity, we do not provide a detailed discussion of empirical research that examines the wide range of other factors potentially related to crime. Wilson and Petersilia (2002) provide an excellent survey of the roles that biology, family, community, schools, labor markets, and intervention programs play in determining crime. Levitt (2004) discusses the impact of the macro-economy, demographics, and gun policy on crime, as well as the evidence that legalized abortion in the 1970s reduced crime in the 1990s because the children who would have been most at risk for crime were never born (Donohue and Levitt, 2001a).

not unified by a single theory of behavior, and they typically use data at the level of individual offenders rather than geographic aggregates. However, like their counterparts in the deterrence literature, these studies are beset by the difficulty of identifying causation and arguably have enjoyed less success in resolving that issue.

Before proceeding, it is worth noting three limits that inhere to social scientific research on the criminal justice system and that this chapter necessarily shares. The first limitation is geographic. The bulk of research on the criminal justice system studies the United States. Second, the chapter primarily concerns the role of publicly administered law enforcement. While the legal system provides opportunities for private enforcement in areas of law such as safety and environmental regulation, civil rights, and antitrust, almost all empirical work on law enforcement focuses on those areas where the government has a monopoly in the use of force and in imposition of incarceration or death as sanctions. Third, social scientific data on the extent and variety of criminal activity pertain almost exclusively to crimes in which victims report the offense to authorities. Because admissions of criminal conduct invite sanctions, self-reports of illegal conduct are inherently unreliable and infrequently collected. Consequently, little empirical evidence exists on illegal activities that are essentially voluntary transactions, such as prostitution or the purchase and consumption of controlled drugs, as well as on crimes in which victims are unaware that they have been harmed, such as insider trading or many frauds. Instead, systematic data are available for—and economists have extensively studied—the traditional categories of criminal activity, such as homicide, rape, robbery, and property crimes. These limitations also constrain the scope of this chapter.

2. Basic theory: economic v. criminologic

2.1. The economic model of crime

2.1.1. Deterrence

The idea that offenders respond to the costs and benefits of crime dates to the eighteenth century, when Beccaria (1770) and Bentham (1789) discussed the concept of deterrence. Becker (1968) provided the first modern and mathematical treatment of the subject, and since his influential article, a large and active literature on the theory of deterrence has developed. Kaplow and Shavell (2002) and Polinsky and Shavell (2007) review comprehensively this literature. Becker's rational choice analysis became synonymous with "the economic model of crime." Because this chapter focuses on the empirical literature on the criminal justice system, we discuss the theoretical literature only briefly to provide the framework for the empirical analysis.

In Becker's model, a criminal rationally maximizes his expected utility. A criminal act causes harm to third parties with certainty, and the offender faces an uncertain punishment. The decision to engage in criminal activity depends on the magnitude of the expected gain from committing the act relative to the expected punishment. If the expected utility exceeds the expected sanction, the individual commits the criminal act.

Several implications of the deterrence model are worth noting. First, more severely punished crimes should occur with less frequency than more lightly sanctioned ones, other things equal. Second, the basic model conceives of the utility-reducing punishment as a monetary fine rather than incarceration. Sanctions in the economic model of deterrence therefore reduce criminal activity through deterrence rather than incapacitation. The next subsection describes the economic analysis of incapacitation. Third, the simple model implies an optimal structure of penalties. Because the expected sanction consists of two components, the probability of apprehension and the magnitude of the punishment imposed on those who are caught, a social planner has two instruments with which to design crime-fighting policies. The probability of apprehension is socially costly because resources must be spent employing a police force. In contrast, fines are socially costless because they are merely monetary transfers. This cost structure implies that for a given level of expected sanction, the optimal penalty structure minimizes the probability of apprehension and maximizes the fine (Becker, 1968; Polinsky and Shavell, 1979). If the model incorporates imprisonment as a sanction, imprisonment is the more costly sanction because the construction and operation of prisons consume resources (Polinsky and Shavell, 1984). Therefore, fines should be used to the fullest possible extent before imprisonment is imposed. A socially costly term of incarceration is not optimal unless the fine is maximal.

The theoretical literature has extended the economic model of deterrence in many directions. An early and influential extension of Becker's framework was Ehrlich's (1973) model of participation in illegal activities as a time-allocation problem. Subsequent models have explored the consequence of relaxing various assumptions of the Becker model. Examples include limited information (Bebchuk and Kaplow, 1992; Kaplow, 1990; Levitt, 1997b; Garoupa, 1999), repeat offending (Polinsky and Rubinfeld, 1991), enforcement error (Png, 1986), and the corruption of law enforcers (Bowles and Garoupa, 1997; Polinsky and Shavell, 2001). Despite the theoretical richness, empirical work on the criminal justice system remains focused on the core aspects of the economic model.

2.1.2. Incapacitation

Economic analysis also offers insights on another purpose of criminal punishment: incapacitation. Unlike deterrence, incapacitation does not pertain to the impact of expected sanctions on the offender's prospective decision to engage in crime. Rather, incapacitation is the reduction in criminal activity achieved by removing the offender from the general population. The goal of incapacitation is not to induce a behavioral response but to restrict the opportunities for the infliction of harm on others.

Shavell (1987) developed a model that evaluated when incapacitation is an optimal policy for an individual offender. In his basic model, incarceration is optimal when the harm done by an offender in a period exceeds the per-period cost of imprisonment. If an offender whose dangerousness exceeds that threshold is constant, the optimal sentence is imprisonment for life, but if criminal propensities decline with age, as is widely observed (Blumstein et al., 1986), incapacitation instructs that the offender should be released when the harm done by an offender falls below the cost of incarceration. In contrast to the optimal structure of sanctions when deterrence is the objective, the efficient sanction for incapacitation purposes is independent of the probability of apprehension.

The predictions of Shavell's (1987) model are broadly consistent with observed criminal justice policies when incapacitation is the exclusive purpose. For example, in *Kansas v. Hendricks*, 521 U.S. 346, 372–74 (1997), the Supreme Court upheld a state law authorizing the civil confinement of sex offenders who have completed their criminal sentences. The Court concluded that further confinement was permissible when offenders suffered from mental abnormalities rendering them likely to re-offend and unlikely to be deterred by the threat of incarceration. Thus, the Court has permitted in limited circumstances confinement solely for the purpose of incapacitation.

Ehrlich (1981) developed the first economic model that distinguished deterrence, incapacitation, and rehabilitation and that emphasized the macro-level implications of incapacitation policies. In particular, he showed that the extent of the "replacement effect," or the degree to which other participants in criminal markets replace incarcerated offenders, are crucial to whether a policy produces reductions in crime. The replacement effect may be large when the supply of new offenders is inelastic, or as Fagan et al. (2003) argue, disproportionate incarceration rates in particular neighborhoods make offending relatively more attractive for remaining residents. Empirical economists have only begun to explore these predictions, and in view of their differing behavioral and policy implications, distinguishing deterrence and incapacitation remains an important challenge for future empirical research on crime.

2.2. *Criminological approaches*

The economic approach to crime differs substantially from the principal schools of criminology.[2] Other than Bentham and Beccaria, early criminologists emphasized the role of biological factors in criminality. Raine (2002) reviews this literature. Cesare Lombroso in his influential *L'uomo Delinquete (The Criminal Man)* described persons who engage in criminal activity as atavistic beings, or reversions to primitive ages (Lombroso, 1876). He claimed that physical characteristics, such as heavy jaws and sloping foreheads, distinguish criminals from non-criminals. Sheldon (1940) developed a more elaborate theory of "somatypes," or body builds that correspond to personality traits. He hypothesized that so-called mesotypes, who possess hard, muscular bodies and aggressive, extroverted personalities, are more prone to criminality. Some criminologists continue to investigate the relevance of immutable characteristics to criminal

[2] This review of criminological theory is merely intended to identify the uniqueness of the economic model of crime. Its explication of the criminology literature is neither complete in detail nor comprehensive in scope.

activity. In their review of the debate between genetic versus environmental explanations of criminality, Shah and Roth (1974) discuss criminological studies of biochemistry, brain and mental disorders, and hormonal imbalances. Researchers have perennially returned to one characteristic, intelligence quotient or IQ. Hirschi and Hindelang (1977) claimed that IQ more accurately predicts delinquency than do race or social class. More recently, Hernnstein and Murray (1994) ignited controversy by emphasizing the correlation between IQ and race. Heckman (1995), one of the many critics of the Herrnstein and Murray study, argued that a single factor does not explain the correlation between test scores and labor market success and that IQ, a supposedly immutable factor, is actually manipulable.

The economic conception of crime, as a rational decision resulting from weighing expected costs and benefits, rejects biological determinism. In the economic model, an individual's endowments or characteristics may affect the attractiveness of criminal opportunities relative to legitimate ones, but no particular characteristic preordains criminal activity. The economic model also assumes that individuals on average respond to incentives. Evidence from the criminological literature on "perceptual deterrence" supports the idea that when individuals perceive a high likelihood of sanction, they are less likely to offend. Nagin (1998) finds that the consensus prevailing in this literature is that perceptions of a risk of sanction correlate negatively with self-reports of offending or of intentions to offend.

Recently, economists interested in the relationship between psychology and economics developed a field of research known as "behavioral economics" that considers how limits on cognitive abilities, willpower, and self-interest affect decision-making. Legal academics, such as Jolls et al. (1998), have begun to explore the implications of cognitive limitations for legal policy. Impairments such as myopia, or shortsighted decision-making, may partly parallel criminological theories about IQ. Whether behavioral economists will draw on criminological research and develop a new behavioral economic theory of crime—and whether such a theory would lend itself to empirical validation—awaits future investigation.

A second major criminological theory stems from Émile Durkheim's concept of "anomie," or an absence of clear norms to direct human behavior. Merton (1968) defined anomie as the consequence of a discrepancy between societal goals and the means available to achieve them. Society places a high value on material success but fails to provide some members of society the instruments to attain it. The individuals experiencing a gap between aspirations and achievement suffer a "strain" and respond by engaging in crime. Cohen (1955) and Cloward and Ohlin (1960) offered their own versions of the strain theory, but both predict that an inability to realize through legitimate channels mainstream or middle class aspirations is a spur to criminal conduct.

This criminological theory yields some predictions that mirror the economic model of crime. In particular, if an individual has preferences similar to other actors but has unattractive opportunities in the legitimate labor market, criminal activity may be a privately optimal choice. On the other hand, strain theory predicts that the most ambitious among those with limited legitimate opportunities will turn to crime, which is not an obvious prediction of the economic model. The limited data on this subject suggest that, indeed, the opposite is true. The economic model further differs from the strain theory in that economics emphasizes individual decision-making rather than social organization as the primary explanation for crime.

A third criminological school considers the social processes through which criminality is learned or culturally transmitted. Sutherland (1956), for example, hypothesized that "differential association," or excessive social contact with advocates of criminal activity, predisposes an individual toward crime. Criminality is thus not the behavior of the physically abnormal, but an activity learned from social interactions with others. Another criminological approach develops theories of social control that analyze how society obtains conformity. Examples include Reckless (1961) and Hirschi (1969). According to Reckless's "containment theory," individuals have social controls or containments that assist them in resisting the social forces pressuring them to commit crime. Under Hirschi's "social bond theory," an individual commits crime when a person's bonds to society, and thus his interest in conforming to society, weaken.

Recently, economists have developed theoretical models containing themes broadly similar to the last two criminological approaches. Some economists relax the standard economic assumption that an individual's preferences are fixed by allowing peer influences to shape them. Sah (1991) posited a model in which an individual's perceptions of the probability of punishment arose endogenously. Glaeser et al. (1996) constructed a model in which the degree of social interaction varies across time, place, and the type of offense. Glaeser and Sacerdote (1999) reported that the higher benefits of urban crime and the lower probabilities of arrest provide only a partial explanation for the higher observed crime rates in cities, and they speculated that a factor perhaps accounting for the remaining difference is the degree of social interaction in cities. Bar-Gill and Harel (2001) developed a model in which crime rates, through their influence on social sanctions, may raise or lower the expected sanction. Just as the social control theorists consider social forces influencing individual behavior, Rasmusen and McAdams (2007) review the emerging law-and-economics literature on social norms.

Overall, the economic model of crime differs from the major branches of criminology in that it abstracts from the social processes and psychological aspects of offending and emphasizes individual choices. A cost of the economic approach is thus a loss of the social context of offending. However, the benefit of its relative simplicity is the set of sharp behavioral predictions that empirical inquiry may validate or refute.

3. Empirical tests of the economic model of crime

3.1. Challenges to empirical testing of the economic model

Despite its conceptual clarity, three obstacles hamper empirical tests of the economic model of crime. First, the available data correspond only roughly to the conceptual model. The next section describes in detail the data used in most economic studies. The economic model of crime is a theory of individual offending, but most empirical analyses examine data aggregated to the city, county, state, or even national levels, which provide only an indirect test of the frequency of individual participation in criminal activity. The typical data are tallies of the number of crimes reported to the police, and as described in the next section, these counts represent a subset of all criminal victimizations (O'Brien, 1985). Worse still, police agencies responsible for official offense counts may manipulate them for political or bureaucratic reasons. Despite these shortcomings, geographic aggregates enjoy a significant advantage over most individual-level data sets of criminal activity: they correspond to the political and legal institutions that set the policies and rules determining expected punishments. This correspondence is crucial because most tests of the economic model compare differences in punishment regimes to geographic variation in crime rates. In contrast, subjects in individual-level data sets of criminal offending are typically drawn from a localized geographic area, and a single punishment regime governs all individuals in a given sample. Observed differences in crime rates are not plausibly attributed to the common punishment regime.

A second major difficulty in testing the economic model of crime is the identification of causal relationships. Simultaneity problems typically beset measurements of the impact of crime-control policies on crime rates. Crime-fighting policies at once set the incentives to engage in crime and respond to the incidence of crime. Specifically, the amount of resources allocated to the criminal justice system is greater when crime is higher. Estimates of the relationship between crime-control policies and crime rates that fail to break this simultaneity, such as ordinary least squares (OLS) estimates from cross-sectional data, are biased. However, the proper identification of causal relationships is a central endeavor of empirical testing of any economic model. Deterrence, the idea that criminal activity responds to expected sanctions, is itself a causal theory. In addition, design of effective public policies to combat crime requires knowledge of causal relationships and estimates of their magnitude.

A third obstacle to testing the economic model of crime is the difficulty of distinguishing incapacitation and deterrence. Incapacitation, unlike deterrence, is not a behavioral response to incentives, but a restriction of the offender's criminal opportunity set. Incapacitation refers to the inability of an imprisoned offender to commit additional crimes as a result of physical constraints on his movement. When criminals tend to re-offend and new entrants into criminal markets do not fully replace them, higher incarceration rates reduce crime through incapacitation, even if the increased expected punishments do not deter. Several surveys show that many prisoners commit multiple offenses when not incarcerated (Chaiken and Chaiken, 1982; Di Iulio and Piehl, 1991), and studies that relate incarceration rates to crime rates thus measure the joint effect of deterrence and incapacitation on criminal activity.

3.2. Data sources

Empirical research on criminal behavior generally relies on three sources of data, each of which possesses its own advantages and shortcomings. The first and oldest data source is the Uniform Crime Reports (UCR). The UCR counts the number of crimes reported to police. Before the UCR, each local law enforcement agency maintained its own figures on the incidence of crime using its own definitions of offenses, and consequently, crime rates between jurisdictions were often not comparable. Begun in 1930, the UCR created a single set of crime definitions for all jurisdictions. Local agencies now report the number of offenses according to the UCR's definitions, and the FBI compiles these tallies and publishes them annually in *Crime in the United States*. Researchers typically aggregate the counts of crimes by geographic units, such as cities or states, and express them as crimes per 100,000 population.

The UCR tracks seven types of crime that comprise the FBI's "Index I" offenses, or simply the Index offenses. The Index consists of four types of violent crime and three types of property crime. The violent offenses, in descending order of severity, are homicide, rape, robbery, and aggravated assault, and property offenses, also in declining order of severity, are burglary, larceny, and motor vehicle theft. The UCR also collects data on arson, but many missing values plague the arson estimates. The UCR's classification scheme is "hierarchical." It counts incidents in which more than one offense is committed as single criminal event. It counts the multiple-offense events as the most severe crime committed during the incident and does not count the other offenses. For example, the UCR would count a "mugging" in which an attacker assaults the victim with a weapon and takes his wallet as a single event, a robbery, because robbery is considered a more severe offense by virtue of its higher position in the UCR hierarchy than assault. The UCR would not count the mugger's assaulting the victim with a weapon as a separate criminal event. The hierarchical structure implies that for offenses other than homicide, the UCR may undercount the actual number of crimes.

The UCR provides no information about the demographics of offenders or victims, or their relationships. An exception is homicide. The UCR has a supplemental file for homicides that contains information on the demographics of offenders and victims, their relationship if any, the type of weapon used, and the time and location of the homicide. Despite their richness, a large number of missing values constrains the utility of the supplemental homicide reports.

The UCR also collects information on arrests. Often researchers divide the number of arrests by the number of offenses to estimate a "clearance rate," or a rate at which police "solve" or clear offenses with arrests. Some researchers claim the clearance rate is a proxy for the probability of apprehension or the efficacy of police, but the arrest data correspond imperfectly to the offense data. The first difficulty is temporal. The UCR arrest data refer to the date of the arrest's occurrence, not the date of the underlying offense. The ratio of annual arrests to offenses does not necessarily reflect police effort at solving the current year's crimes. In fact, because the number of arrests in a year may exceed the number of reported offenses, the annual clearance rate may exceed unity. Second, a low clearance rate does not necessarily imply the inefficacy of police, because a repeat offender may be responsible for many reported crimes, but his apprehension is recorded as only one arrest. The arrest data encounter another difficulty when they are decomposed by offense category. The fact that roughly 80% of arrests are for offenses other than Index crimes (FBI, 2002) suggests that the offense for which an individual is arrested may not refer to the actual crimes committed for several reasons. Police might arrest a person for a lesser offense if they lack sufficient evidence to prosecute him for the most serious crime of which they suspect him. Or, they may arrest a person on a minor crime, such as a public order offense, and after investigation, learn that the arrestee is suspected of more serious crimes. This latter possibility seems increasingly likely as the number of unserved felony arrest warrants grows and the comprehensiveness of DNA data bases expands.

Despite their shortcomings, the UCR's Index offense counts are the data that empirical economists use most widely. The advantages of the UCR are its consistency and geographic disaggregation. The uniformity and constancy of the UCR's offense definitions permits comparisons of crime rates across jurisdictions and across time. The geographic disaggregation of the UCR permits linking the crime data to other geographically-defined data that vary across time and jurisdiction, such as labor market outcomes, population demographics, and criminal justice policies. The UCR thus lend themselves to panel estimation of the relationships between the crime rates and crime policy variables. Other data sources provide greater detail on criminal events but are less useful because they lack these temporal and geographic dimensions.

The second major source of data on criminal activity is the National Crime Victimization Survey (NCVS), or as it was known prior to 1990, the National Crime Survey (NCS). The NCVS is a survey of households that gathers information on the incidence of criminal victimization. It collects information on the demographics of offenders and victims and the circumstances of their victimization. It contains information on the same offense categories as the UCR does, except for homicide (there are no victims to respond) and arson. The U.S. Census Bureau conducts the survey and publishes the results annually in *Criminal Victimization in the United States*.

The NCVS regularly implies a crime rate higher than the UCR data. Whether this discrepancy reflects the greater accuracy of the NCVS is unclear. Respondents to a survey may be more willing to report a crime that they were too embarrassed to tell the police or that was too small to warrant the hassle of informing the police. Alternatively, the NCVS may over-report the incidence of offenses because respondents "telescope," or harken back to distant offenses, rather than describe only recent victimizations.

For empirical economists, a significant limitation of the NCVS is its aggregation at the national level. In order to maintain the anonymity of the respondents, the NCVS data are not decomposed by geographic unit. Empirical researchers thus cannot link the NCVS data to variation in laws and crime policies across jurisdictions. This constraint renders the NCVS of little use in assessing the impact of such policies on criminal activity.

The U.S. Department of Justice is currently developing another data source that combines high degree of detail of the NCVS with the geographic decomposition of the UCR. The National Incident-Based Reporting System (NIBRS) gathers information from police reports about the circumstances of and participants in a criminal incident. Rather than applying the hierarchical crime-counting scheme of the UCR, the NCVS reports each offense separately. The NBIRS may thus combine the advantages of the UCR and the NCVS while shedding many of their disadvantages. Presently, only a handful of jurisdictions report under the NBIRS, and this limited coverage stymies wider use of the NBIRS by researchers. If NBIRS reporting expands, it will be an important data source for future empirical research.

The third major source of data on criminal activity relies on self-reports of offending, rather than self-reports of victimization. The National Longitudinal Survey of Youth (NLSY) contains a nationally representative sample of youth and periodically inquires about respondents' participation in criminal activity. The data are individual-level observations rather than geographic aggregates. The types of offenses that the NLYS inquires about naturally tend to be less severe, such as drug taking. However, the data allow direct measurement of these activities as well as demographic and economic factors. Empirical economists have used the NLSY in several important studies of the relationship between criminal offending and schooling and the labor market.

3.3. Testing the economic model using the scale of policing

Police are perhaps the most visible feature of the criminal justice system, and academics have long studied the relationship between the scale of policing and the level of criminal activity. In the 1970's numerous studies attempted to test this relationship. The most prominent of these was the quasi-randomized experiment in Kansas City (Kelling et al., 1974). For one year, three areas of the city, each consisting of five police beats, received different levels of policing. In the first, the level of policing was the standard amount: police responded to citizens' calls for service and one car engaged in preventative patrols. In the second area, the level of policing was greater than the standard: in addition to responding to calls for service, two to three police cars regularly patrolled. In the last, the policing level was below the standard: no police vehicles patrolled and police only responded to service calls. The authors considered various outcome measures, from victim reports to public perceptions, but few statistically significant differences between the three areas were seen. Larson (1976) criticized the appropriateness of the conclusion that the experiment had no effect on outcomes. Larson claimed that because response times to calls for service in the three areas exhibited no change, the experiment failed to induce sufficient variation between treatment and control groups to warrant any conclusion.

Subsequent studies did not intentionally vary the level of police protection. Instead, numerous studies compared cross-sectional differences in police and crime at a point in time. Cameron (1988) found that 18 of the 22 such studies he reviewed reported either no relationship or a positive relationship between the level of police and crime rates. This finding appears to refute the economic model of crime, but Fisher and Nagin (1978) persuasively showed that these studies, by drawing cross-sectional comparisons, suffered from simultaneity bias. Jurisdictions with higher crime rates react by hiring more police, and this response produces a positive cross-sectional correlation between police and crime rates. The correlation, however, reveals little about whether the additional police cause reductions in the rate of crime. For this reason, a report by the National Academy of Sciences (NAS), which included the Fisher and Nagin (1978) study, issued a blistering critique of the cross-sectional methodology. Researchers largely abandoned the subject of police and crime for the next two decades.[3]

Beginning in the late 1990's, the literature on policing and crime rates experienced a renaissance. This second wave of literature enjoyed two principal advantages over earlier studies: the use of richer and larger data sets and the application of more sophisticated econometric techniques. All of the new research moved beyond single cross-sectional data sets and instead examined repeated cross-sections that often spanned more than two decades. Larger data sets with more observations improve the precision of the estimates. More importantly, repeated cross-sections permit identification of coefficients from variation in crime rates within a jurisdiction over time rather than merely across jurisdictions. Unobserved heterogeneity is less likely to characterize within jurisdiction variation in crime rates.

The second innovation of the new literature on crime and policing levels was the application of novel strategies to break the simultaneity problem. Each of these studies employed a different strategy, and remarkably, each obtained similar results. These studies found that higher policing levels cause large reductions in crime rates.

The first of these studies, Marvell and Moody (1996), estimated the impact of policing levels on crime rates using the technique of Granger causality. One variable "Granger causes" another when changes in the first variable generally precede changes in the second, and thus, Granger causality refers to a temporal relationship between two variables rather than actual causation (Granger, 1969).[4] Specifically, Marvell and Moody (1996) estimated two autoregressive equations:

$$C_{i,t} = \sum_{j=0}^{J} P_{i,t-j}\delta_j + \sum_{j=0}^{J} C_{i,t-j}\alpha_j + X_{i,t}\gamma_1 + \varepsilon_{i,t} \tag{1}$$

$$P_{i,t} = \sum_{j=0}^{J} P_{i,t-j}\phi_j + \sum_{j=0}^{J} C_{i,t-j}\varphi_j + X_{i,t}\gamma_2 + \upsilon_{i,t}. \tag{2}$$

Equation (1) is a regression of the (natural logarithm of) the crime rate per 100,000 persons in jurisdiction i in year t, $C_{i,t}$. It is modeled as a function of J lags of itself

[3] Researchers' unwillingness to pursue this topic in wake of the severity of the NAS's rebuke might itself be evidence of deterrence.

[4] Students of rhetoric recognize Granger causality as the "post hoc, ergo propter hoc" (after this, therefore because of this) logical fallacy. A frequently-cited example of the difference between actual causation and Granger causality is the delivery of Christmas cards (Kennedy, 1998). Christmas cards arrive in mailboxes before December 25, and thus, they Granger cause Christmas. But, the true causation is the reverse: the imminent arrival of Christmas causes individuals to mail Christmas cards. Forward-looking behavior produces a divergence between actual causation and Granger causation. Marvell and Moody's evidence likely indicates true causation because forward thinking behavior is not a plausible alternative explanation in the context of aggregate crime rates and levels of policing.

and J lags of $P_{i,t}$, the (natural logarithm) of police officers per 100,000 persons in jurisdiction i in year t. If the coefficients on the lags of the policing levels, δ_j, are jointly significant, changes in the size of the police force Granger cause movements in crime rates. Equation (2) is an analogous regression in which policing levels are a function of lags of itself and of crime rates. If the coefficients on the lags of the crime rates in this equation, α_j, are jointly significant, crime rates Granger cause policing levels. The matrix $X_{i,t}$ contained other explanatory variables specific to each jurisdiction and year. The inclusion of jurisdiction-specific time trends further reduces the likelihood that omitted variables biased the coefficients of interest. Estimating these equations on more then twenty years of state and city data, Marvell and Moody (1996) found that increases in police were associated with future declines in crime. That is, police Granger caused lower crime. These authors showed that estimation strategies that account for the simultaneity yield substantially different conclusions about the relationship between police and crime. They suggest that increases in the size of police forces depresses crime rates.

Levitt (1997a) employed a different methodology, instrumental variables, but also found that larger police forces associate with declines in crime rates. He estimated a system of equations of the form:

$$\Delta P_{i,t} = \Delta Z_{i,t}\Theta + \Delta X_{i,t}\beta_1 + \nu_{i,t} \tag{3}$$

$$\Delta C_{i,t} = \Delta P_{i,t}\Phi + \Delta X_{i,t}\beta_2 + \omega_{i,t}. \tag{4}$$

As in equations (1) and (2), $C_{i,t}$, $P_{i,t}$, and $X_{i,t}$ represent (natural logarithms) of police rates, crime rates, and a matrix of control variables, respectively. Unlike Marvell and Moody (1996), who examined levels, Levitt (1997a) expressed these variables as first-differences. First-differencing revealed a negative but weak relationship between police and crime and suggested that unobserved heterogeneity across states biased OLS estimates upward.

As the equations show, a valid instrument, $Z_{i,t}$, correlates with police but does not correlate with crime, except through the other explanatory variables in the crime equation. Levitt (1997a) argued that mayoral and gubernatorial elections satisfied this identifying restriction. He showed that sizable increases in the police forces in major cities occurred in election years, presumably because this timing generates electoral benefits for politicians. After controlling for other factors, this instrument was arguably exogenous to crime rates, because the occurrence of an election is unlikely to affect incentives for criminal activity, other than through these changes in the police force. The elections were effectively a "natural experiment" that induced movement in the size of police forces but were otherwise exogenous to crime rates. Estimating equations (3) and (4) using two-stage least squares (2SLS), Levitt found that additional police induced sizable decreases in crime rates. The instrumented estimate of Φ implied that an increase of 10% in the police force lead to an estimated 3 to 10% reduction in crime rates. These estimates were comparable to those of Marvell and Moody (1996). McCrary (2002) disputed that electoral cycles induced sufficient variation in the size of police forces to measure the impact of crime. He argued that when properly measured, the imprecision of the instrument precluded inferences about the causal effect of police on crime. In response, Levitt (2002) showed that even with an alternative instrumental variable, the number of firefighters, estimates of the impact of police on crime are negative and sizable. More recently, Evans and Owens (2004) have shown similar impacts on crime of increases in police hired with federal subsidies accompanying the Clinton Crime Bill.

A third approach is the use of more finely disaggregated data. Corman and Mocan (2000) demonstrated the advantage of using higher frequency data. Prior studies used annual observations of police and crime, but these authors showed that police hiring occurs approximately six months after a surge in crime. An increase in crime at the beginning of a year would therefore result in an increase in police later in the year. With annual observations, the increase in crime and police would appear contemporaneous rather than sequential, and the short-term causal effect of police on crime would not be observed. Corman and Mocan used almost thirty years of monthly data from New York City to capture these short-term changes and test for Granger causation. Their estimates indicated that a 10% increase in the police force produced a 10% reduction in crime rates.

Analogously, Di Tella and Schargrodsky (2004) tested for the crime-reducing effect of police using data with a fine geographic decomposition: at the level of city blocks. They examined the impact on crime of a dramatic police re-allocation in Buenos Aires in response to terrorist threats against temples and mosques. Police presence outside these buildings increased sharply. These authors found that auto theft on the blocks on which temples and mosques were located fell precipitously, but the crime reduction quickly decayed with distance. This paper confirmed that the presence of police deters crime, although the broader applicability of the findings may be quite limited.

Based on these estimates of the impact of police on crime, it appears that the marginal benefit in reduced crime associated with hiring an additional police officer in large urban environments (where the benefits are likely to be greatest) are roughly in line with the marginal cost. Thus, policy makers have seemingly managed to achieve a scale of policing that appears approximately socially efficient. Although economists' understanding of the causal relationship between policing and crime has improved, they have made less progress on the question whether the declines in crime attributable to increased police are due to deterrence or incapacitation. More police could deter criminals from committing crimes because greater numbers of police reduce the probability they escape detection and punishment. Alternatively, more police could have no deterrent effect. They could simply raise the probability that repeat offenders are caught and imprisoned, thereby incapacitating their commission of additional offenses.

While the economic model of crime predicts that more police reduce *crime* rates, it predicts an ambiguous effect on *arrest* rates. More police could produce higher arrest rates because the probability of detection rises. Alternatively, more police could induce a behavioral response from offenders. Knowing that the additional police raised the risk of apprehension, criminals could undertake fewer offenses. The only study to examine this question in an empirically convincing manner does so in the context of sports.

McCormick and Tollison (1984) studied the effect of increasing the number of college basketball referees from two to three. In their analogy, referees correspond to police and fouls to arrests. Their estimates showed that the addition of a referee correlated with a 30% reduction in the number of fouls. Although the estimated impact is considerable, whether it generalizes to the context of crime is unknown.

3.4. Testing the economic model using the scale of imprisonment

3.4.1. Estimates of the crime-reducing effect of the scale of imprisonment

Prisons are another highly visible and widely used crime-control policy, and research on the crime-reducing effect of prisons has followed a similar path as studies of policing. Early efforts failed to recognize or address the simultaneity problem and concluded that imprisonment has no impact on crime rates. More recent studies identified causation through Granger causality and instrumental variables techniques and reached strikingly different conclusions. The early literature on the scale of imprisonment compared only time-series variations in national rates of incarceration and crime. Prison populations grew substantially in the 1970's, and this growth accelerated in the 1980's. Crime rates, however, continued to rise until the 1990's. Zimring and Hawkins (1991) interpreted the pattern of rising national incarceration and crime rates as evidence that imprisonment had neither deterrent nor incapacitating effects. Concluding that imprisonment was a failed public policy, they called for a reduced use of incarceration. The shortcoming of this conclusion is its failure to distinguish correlation and causation. Like early work on the size of police forces, the scale of imprisonment is partly a function of political forces that increase prison populations incarceration rates in response to higher crime.

Beginning in the mid-1990's, several researchers applied more sophisticated empirical strategies that attempted to disentangle correlation. Marvell and Moody (1994) examined time-series patterns in crime and imprisonment at the state rather than national level and again applied the techniques of Granger causality. They estimated regressions analogous to equations (1) and (2), only with (lags of) incarceration rates replacing the policing variables, $P_{i,t}$. Marvell and Moody (1994) first observed that cross-sectional comparisons of states at a point in time showed a positive association between prisons and crime. However, within a state over time, incarceration rates did not exhibit a positive correlation with crime rates. Applying Granger-causation techniques, Marvell and Moody found that a 10% increase in the prison population produced a 1.5% fall in crime rates.

Levitt (1996) provided further evidence of the crime-reducing effect of imprisonment. As in his research on police, he used an instrumental variable to break the simultaneity of crime and incarceration. Levitt (1996) contended that lawsuits sponsored by the American Civil Liberties Union (ACLU) and alleging overcrowded prisons violated inmates' constitutional rights induced variation in prison populations. The lawsuits were arguably a valid instrument, because these reductions are exogenous to crime rates. He showed that when the ACLU suits produced court orders to reduce overcrowding, states typically complied by releasing prisoners who would otherwise have been incarcerated. He specified a system of equations analogous to equations (3) and (4), only with prison populations replacing the police staffing variables. His 2SLS estimation treated prison populations as endogenous and the other explanatory variables in the matrix $X_{i,t}$ as exogenous. His analysis showed that when states released prisoners as a result of the litigation, they experienced a coincident uptick in crime rates relative to other states. His estimates implied that the reduction in crime from incarcerating an additional prisoner was two to three times larger than that predicted by Marvell and Moody (1994). According to Levitt's (1996) analysis, the release of an additional prisoner produced an additional 15 crimes annually. The costs of incarcerating an offender are high. Recent estimates of the annual cost range from \$25,000 (Piehl and DiIulio, 1995) to \$35,000 (Donohue and Siegelman, 1998). The attractiveness of incarcerating the marginal prisoner from a cost-benefit perspective depends critically on the estimate of the impact of imprisonment on crime, as well as on the assumptions made regarding the rate at which the social harms caused by the marginal prisoner decrease. There are strong reasons to believe that the two millionth prisoner incarcerated is far less dangerous than the first prisoner incarcerated. Because the prison population has grown so quickly, evaluation at the margin of the current scale of incarceration takes one well outside the range in which the parameter estimates were generated. For most reasonable sets of assumptions, it appears that the current scale of incarceration is above the socially optimal level.

3.4.2. Distinguishing deterrence and incapacitation

The studies described in the previous section provide compelling evidence that increases in prison populations cause declines in crime rates. They do not, however, distinguish whether deterrence or incapacitation is responsible for the reductions in crime. Determining the operative effect is important for testing the economic model of human behavior. In addition, as described earlier, the efficiency of a crime-control policy depends crucially upon whether the goal is deterrence or incapacitation.

A perhaps direct but econometrically unsophisticated approach to investigating the relative importance of deterrence is to ask criminals (and potential criminals) if threatened sanctions influence their behavior. Such surveys may be inaccurate if respondents intentionally mislead or lack self-awareness of their decision-making processes. Despite these drawbacks, surveys have found evidence generally consistent with deterrence, but the effect is weak and often not the most important factor (see, for example, Saltzman et al., 1982). However, the effect is even found among juveniles, a demographic group often thought to be less forward-thinking in its decision-making. Glassner et al. (1983) found a sizable fall in self-reports of criminality at the age of majority, when respondents became eligible for the harsher sanctions of the adult criminal justice system. Self-reports thus suggest that incapacitation is not wholly responsible for the crime-reducing effect of incarceration.

A more systematic examination of the relative importance of deterrence and incapacitation tests differences in the behavioral implications of deterrence and incapacitation.

Levitt (1998a) predicted that if the sanction for one particular type of crime rises, then deterrence implies that generalist criminals should substitute to other kinds of offenses. For example, if the punishment for robbery rises, deterrence predicts that criminals switch to commit more burglaries. If instead the primary effect of a longer sentence is incapacitation, the rate of all substitute offenses should decline as well. For example, if the length of sentences for robbery rises, incapacitation would reduce the rates of both burglary and robbery. Evidence indicates that criminals do not specialize in a particular type of offense, such as robbery, but instead are generalists who participate in potentially wide range of offenses (Beck, 1989). Using data on arrest rates, Levitt (1998a) tested for substitution between types of criminal activity as a means to assess the relative importance of deterrence and incapacitation. He found that for certain crimes, such as rape, the incapacitation effect dominated and for other crimes, such as property offenses, deterrence explained 75% of the observed reduction in crime. For still other offense categories, such as robbery, the effects of deterrence and incapacitation were roughly equal in magnitude.

A question related to the degree of substitution between offenses is whether prison sentences are efficiently allocated among different kinds of offenders. A controversial aspect of the expansion of incarceration is the fifteen fold increase in the number of imprisoned drug offenders. A widely held (though rarely tested) view is that incarcerating drug offenders prevents few other crimes because their illegal activities are limited to drug trafficking or consumption. Kuziemko and Levitt (2004) found evidence refuting this hypothesis. Their estimates indicate that incarcerating a drug offender reduces the expected time served by inmates convicted of less serious offenses. However, incarcerating a drug offender reduces violent and property crime almost as much as incarcerating any other type of offender.

Other efforts to assess the relative magnitudes of deterrence and incapacitation have examined the predicted impact of changes in the laws governing punishments on criminal behavior. These legal variation permit use of difference-in-difference estimation. For example, Shepherd (2002) studied states' adoption of truth-in-sentencing laws that required violent offenders to serve 85% of their sentences. She found, consistent with Levitt (1998a), that after a state adopted a truth-in-sentencing law, the rate of violent crime dropped and the rate of property crime rose, relative to other states.

Kessler and Levitt (1999) used a different legal change to obtain another assessment of the relative significance of deterrence and incapacitation. The authors examined the effect of a referendum in 1982 in California, known as Proposition 8, that provided sentence enhancements for certain crimes. Before the adoption of Proposition 8, the affected offenses were serious crimes for which a conviction almost always garnered a prison term. The sentence enhancement therefore had an incapacitating effect only upon completion of the standard prison term, and any decline in crime before that date was attributable to deterrence. Kessler and Levitt used a "differences-in-differences-in-differences" strategy that compared eligible offenses to ineligible offenses, before and after the passage of Proposition 8, in California and the rest of the country. They found that after the passage of Proposition 8, the rate of eligible offenses in California fell more than the rate of non-eligible offenses, while in the rest of the country, the movement in the rate of eligible crimes relative to ineligible crimes was more modest. Their estimates indicated that the sentence enhancements reduced eligible crimes by 4% in the year after adoption and by 8% in the three years after adoption. Their results suggest that deterrence and incapacitation may each explain about half of the reduction in crime attributable to the sentence enhancements.

Levitt (1998b) offered another means of distinguishing deterrence and incapacitation. Rather than comparing differences in behavior before and after the introduction of more severe punishments, he evaluated the responsiveness of criminal activity to the transition from the juvenile to the adult criminal justice system. In states where the criminal justice system is substantially more punitive than the juvenile system, deterrence predicts that juveniles should reduce their criminal activity upon reaching the age of majority. Levitt (1998b) found that in states where the adult system is significantly more punitive than the juvenile system, violent crimes fall by nearly 4% at the age of majority, and where the transition to the adult system is most lenient, violent crime rises by more than 23%. For property crimes, these figures are roughly a 20% decline where the transition is harsh and a 10% increase where it is not. The estimates were consistent with the survey evidence of Glassner et al. (1983) in that they show more severe sanctions deter even juveniles.

The distinction between deterrence and incapacitation is particularly crucial for policies, such as "three strikes laws and you're out," that sentence repeat offenders to life in prison without parole. If incapacitation is the primary effect of incarceration, then policies of permanent incarceration are likely inefficient. The declining age-crime profile, the phenomenon that participation in criminal activity falls with age (Blumstein et al., 1986), implies that inmates serving life sentences may remain incarcerated long after their risk of recidivism becomes negligible (Shavell, 1987). The costs of incarcerating aged criminals would surely exceed the social benefits of their averted crimes. If instead deterrence is the primary effect of incarceration, severe sentences could be efficient. Long sentences could dissuade potential criminals from participating in illegal activities, and with fewer offenses committed prison populations could actually fall.

During the 1990's, a number of state legislatures passed three-strikes laws, and the experience of California, the one state that meted out third strike sentences with regularity, provided some anecdotal evidence of deterrence. Greenwood et al. (1994) predicted that if the three-strikes law were fully implemented in California, annual prison expenditures would nearly triple. However, between 1994 and 1998, California's prison population grew at 29%, only 6 percentage points more than the national average, and its violent crime rate fell 30%, which was 10 percentage points greater than the national average. These patterns are consistent with the presence of some deterrent effect, but a formal analysis is necessary.

Economists have made substantial progress in disentangling causational issues plaguing the tests of whether prisons reduce crime. Attention has shifted from whether prisons reduce crime to understanding the mechanism of how prisons reduce crime. The evidence to-date suggests that incarceration may have both substantial incapacitation and deterrent effects, but their relative importance may vary by type of offense and type of offender. Despite the empirical challenges of identifying the operative effect, more research is needed on this important question.

3.5. Testing the economic model using capital punishment

Capital punishment is perhaps the most controversial criminal justice policy, and it has long been debated both within academia and among the general public. For the empirical researcher, capital punishment seemingly offers an ideal test of deterrence. The alternative sentence for an offender who is eligible for the death penalty is typically permanent incarceration. Any behavioral responses to the imposition of the death penalty rather than a life sentence are therefore arguably due to the deterrent effect of capital punishment rather than incapacitation. Although the death penalty appears at a conceptual level to provide a direct test of the deterrence model, the actual manner of its implementation complicates empirical testing.

The researcher most closely identified with empirical evidence of capital punishment's deterrent effect is Isaac Ehrlich. As research on the death penalty evolved from time-series studies, to cross-sectional studies, and finally to repeated cross-section analyses, Ehrlich found with each methodology estimates of a large deterrent effect of capital punishment. Numerous other researchers assailed Ehrlich's conclusions, and Cameron (1994) provides an extensive survey of the large empirical literature on capital punishment.

The first generation of studies on the deterrent effect of the death penalty examined national time-series data. Ehrlich (1975) estimated that over the period 1932–70 each execution deterred between one and eight homicides. His results also show sizable reductions in robbery, aggravated assault, and property crimes. Other researchers asserted that if the sample period is extended to include later dates or truncated to exclude the late 1960's, no effect on crime rates is found (Passell and Taylor, 1977; Klein et al., 1978). For academic studies, the estimates received an usual degree of public attention. Even the Supreme Court, in *Gregg v. Georgia*, 428 U.S. 153 (1976), which held that capital punishment could comport with the Eighth Amendment, commented on the debate. In reference to the estimates on the deterrent effect, the Court observed, "[s]ubstantial attempts to evaluate the worth of the death penalty as a deterrent to crimes by potential offenders have occasioned a great deal of debate. The results have simply been inconclusive." *Gregg*, 428 U.S. at 184–85.

Similar controversy plagued cross-sectional estimates. Using cross-sectional data for states in 1940 and 1950, Ehrlich (1977) found results close to those of his time-series study. As discussed above in the context of policing and incarceration, cross-sectional estimates may not reliably identify a causal effect, because a particular state's criminal justice policies, including the use of capital punishment, may itself be a function of a state's crime rate. The concentration of executions in Southern and Western states strongly suggests that the use of the death penalty is non-random. Cameron (1994) again summarizes particular criticisms of Ehrlich's econometric approach. Some researchers lodge methodological criticisms. Leamer (1983) and McManus (1985) apply extreme bounds analysis and report that cross-sectional estimates of the effect of executions on the homicide rate are highly sensitive to the choice of control variables. Ehrlich and Liu (1999) criticize extreme bounds analysis for risking rejection of valid models and acceptance of invalid ones. They argue that economic theory should guide the choice of control variables, and they present estimates from a repeated-cross section consistent with a deterrent effect. Other researchers examine other predictions of the deterrence hypothesis. For example, Bailey (1982) reports no effect of the death penalty on the frequency with which police officers are killed. Because killers of police officers are much more likely to face capital punishment than murderers of civilians are, the absence of an effect there is troubling.

Like the results from the early national time-series studies, estimates from repeated cross-sectional data are sensitive to the time-period considered. Katz et al. (2003) used state-level panel data covering the 1950–90 period and detected no effect of the death penalty on crime rates. Katz et al. (2003) found instead that a proxy for prison conditions, the death rate among prisoners, correlated negatively with crime rates. In contrast, studies that consider the period since 1976, the year of the Supreme Court's decision in *Gregg* that reinstated the death penalty as a constitutionally permissible punishment, report large deterrent effects. Mocan and Gittings (2003) considered the effect of commutations, as well as executions, in a monthly, state-level panel spanning 1977–97. They reported that the marginal execution saves five homicides, and each commutation causes five homicides. Their estimates indicated that executions and commutations had no effect on other categories of crime, such as assaults, robberies, or property offenses. Dezhbakhsh et al. (2003) used annual, county-level data in a simultaneous equations approach and claimed an enormous deterrent effect. They assert that, with a margin of error of ten, each execution prevents 18 homicides; the lower end of their estimated impact, eight homicides, is equivalent to the highest of Ehrlich's (1975) estimate. Shepherd (2004) used a monthly, state-level panel covering 1977–99 to estimate the effect of the death penalty on different types of homicides. Her estimates indicated that each execution deterred approximately three murders, including so-called "crimes of passions," and that the execution deterred homicides by both whites and African-Americans. Shepherd also estimated the deterrent effect of the length of time that condemned prisoners spend waiting their executions on death row. She reported that each 2.75-year reduction in the death row queue deterred one homicide. In sum, even when panel data are used, evidence of capital punishment's deterrent effect remains sensitive to minor specification changes. Unlike the literature on policing and incarceration, the use of higher frequency data and additional control variables has broadened, rather than narrowed, the range of estimated impacts. Donohue and Wolfers (2006) review and re-analyze the studies covering the post-moratorium period. They find that the estimates are sensitive to modest changes in specification and doubt whether existing data on the death penalty have sufficient variation to support inferences about its impact.

A large deterrent effect is surprising given the relatively abstemious application of the death penalty. In 1999, 98 prisoners were executed, more than any other year since the 1976 reinstatement of capital punishment in *Gregg*. However, relative to the 3,540 prisoners under a sentence of death in 1999, the number of executions constituted 3% of the death row. Since then, the number of death row inmates has steadily grown while the pace of executions has slowed. For example, the 66 executions in 2001 implied an execution rate of 1.8%. The execution rates on death row are only slightly larger than the risk of accidental and violent death for African-American males aged 15 to 34. The new death sentences as a percentage of the homicides in particular year, a rough proxy for the risk of receiving a death sentence conditional on having committed a homicide, is also low—hovering between 0.9% and 1.8% in recent years. Note also that most criminals sentenced to death will never actually be executed, so this number dramatically overstates the true risk of execution. For a person regularly participating in crime, the risk of violent death is surely higher. Kennedy et al. (1996) estimate that annual violent death rates among gang members in Boston are between 1% and 2%. Levitt and Venkatesh (2000) find that the death rate among street-level drug dealers in an urban gang was 7%. Individuals who regularly participate in criminal activities with such hazards are unlikely to be influenced by the relatively low risk of capital punishment. Ironically, to the extent the death penalty has any deterrent effect on homicide, it is unlikely to be the rational criminal who is deterred, but rather, an irrational criminal who mistakenly overstates the expected punishment associated with the death penalty.

3.6. Testing the economic model using victim precautions

As the foregoing discussion indicates, most empirical economic research on deterrence investigates the effect of variations in the criminal justice system. A smaller number of studies test the responsiveness of crime to the precautions taken by potential victims. Like government expenditures on law enforcement, private precautions against crime raise the expected cost of committing crimes, but unlike public law enforcement, private precautions do not raise the expected cost of committing crimes uniformly across all potential victims. The nature of a criminal's response to victim precautions depends on whether he can observe the precautions before deciding to commit the offense (Clotfelter, 1978; Shavell, 1991). If the offender observes a precaution, he may forgo the protected target in favor of an unprotected one. Observable precautions thus deter the commission of crimes against specific victims, those with hardened targets. Observable precautions, however, do not reduce the aggregate amount of crime when offenders can readily substitute to unprotected victims. If, instead, an offender cannot observe whether a potential victim has taken a precaution, he cannot substitute to an unprotected victim. An unobservable precaution thus raises the expected cost of offending for all potential victims, and thus, may deter the commission of crimes generally.

Recently, a series of studies of state laws permitting citizens to carry concealed weapons sparked enormous controversy. Lott and Mustard (1997), Lott (1998), and Lott and Landes (1999) claim that a victim's surreptitious gun possession is an unobservable precaution that raises the cost of criminal conduct and thus has a general deterrent effect. Lott and his co-authors compared county-level crime rates before and after a state's passage of concealed weapons laws, relative to other states without these laws. They claimed that their difference-in-differences estimates showed that following the enactment of these laws, the rates of certain types of crime decline substantially.

The Lott studies constituted a sharp challenge to the existing literature on guns and crime, which focuses on the negative aspects of guns on public health. Cook et al. (2002) provide a thorough review that literature. Following publication of the original Lott and Mustard (1997) study, numerous researchers challenged their findings by exploring the robustness of their results and conducting additional tests of their unobservable precaution hypothesis. Duggan (2000) showed that the Lott and Mustard (1997) estimates for robbery and larceny lack statistical significance when the assumption of the statistical independence of counties within the same state is relaxed. Duggan also demonstrated that the passage of a concealed weapons law did not correlate with a proxy for the rate of gun ownership. The Lott estimates also appeared to suffer from omitted variable bias. Donohue and Levitt (2001a) report that after controlling for abortion rates, the laws did not correlate with crime rates. Ayres and Donohue (2003) extended the original Lott and Mustard (1997) sample to encompass seven additional years of data, including more recent enactments of concealed weapons laws, and in this larger sample, they found no negative correlations between the law and crime rates. Further tests of the behavioral implications of the concealed-weapons hypothesis have also raised questions about its validity. For example, the more frequent possession of concealed weapons should have the greatest deterrent effect on crimes involving the criminal's direct confrontation of a stranger victim. Street robbery is the prime example of such a crime. But, the original Lott and Mustard (1997) estimates featured large effects on homicide and rape and a much smaller effect on robbery. Moreover, Ayres and Donohue (2003) reported that in their extended sample, the concealed weapons laws correlated positively with robbery rates. Ludwig (1998) observed that only adults are eligible for concealed weapons permits, and he predicted that any additional deterrence from the laws should occur primarily in adult populations. He employed a "differences-in-differences-in-differences" estimation framework in which juveniles were the control group and adults were the treatment group. Ludwig's estimates showed that contrary to the Lott hypothesis, concealed weapons laws have a weak and even positive effect on the rate of adult homicides relative to the rates for juveniles.

The other principal empirical study of unobservable precautions is Ayres and Levitt's (1998) analysis of Lojack. Lojack is an anti-theft device for automobiles. Once hidden in a vehicle, the device contains a radio-transmitter that allows police with special equipment to track the vehicle. Unlike most car alarms that emit loud sounds when the car is disturbed or that come with window stickers announcing the alarm's installation, Lojack does not betray its presence with outward indications, even after the car is stolen. Lojack is therefore an ideal unobservable precaution. Ayres and Levitt (1998) found that when Lojack becomes available in a city, vehicle thefts fall sharply. Lojack facilitates the apprehension of car thieves by leading police to so-called chop shops, where stolen vehicles are disassembled for parts. The authors' estimates show that Lojack does not induce car thieves to substitute to other types of crimes or to other geographic areas.

The results suggest that the owners of cars with Lojack capture only part of its deterrent benefit, and therefore, the use of Lojack may lie below the social optimum. Despite enormous private expenditures on crime prevention, studies on unobservable precautions are relatively scarce, and reliable estimates of the degree of substitution induced by observable precautions are non-existent. More research on the relationship between private precautions and crime is needed.

4. Empirical study of particular aspects of the criminal enforcement system

4.1. Challenges to empirical study of the criminal enforcement system

Although tests of the economic model of crime are the economists' primary contribution to the empirical study of the criminal justice system, economists have also examined particular institutions within the criminal justice system. For several reasons, the empirical literature on the criminal justice system is not as well developed as tests of the economic model of crime. First, a single conceptual model does not unify empirical studies of criminal justice institutions in the same way that Becker's model of deterrence guides the economists' study of criminal behavior. The Beckerian framework is a model of offending behavior that largely abstracts from the criminal justice system. Legal institutions are relevant to the economic model to the extent that they influence the expected penalty through the probability of punishment or the magnitude of the penalty.

In contrast, empirical studies of particular criminal justice institutions typically do not test the deterrence hypothesis. They instead consider the operation of specific legal institutions, actors, or procedures. In lieu of a single unifying conceptual model, authors of these studies typically begin by positing a behavioral framework specific to the particular inquiry. The models often lack wider application and are tailored to the particular data set that a researcher possesses. Second, a lack of readily available data hampers economic studies of the criminal justice system. Unlike crime rates, no government agency produces annual, national data on criminal prosecutions, sentences, and appeals in state courts. Although the U.S. Sentencing Commission makes available data on the sentences of federal defendants, federal policies lack the natural cross-sectional variation of state policies, and the absence of cross-state variation inhibits identification of causal effects. Economists interested in empirical analysis of state criminal courts must therefore engage in costly data collection. Even when studies address similar research questions, their use of different data sets limits the comparability of their results. Third, the criminal law and criminal procedure are large, complex, and continually changing bodies of knowledge requiring considerable expertise to master. Similarly, the criminal justice system is an elaborate institution involving many actors and procedural nuances. Economists generally do not possess a sufficiently sophisticated understanding of these laws and institutions to ask—and obtain persuasive answers to—questions of importance to legal scholars and practitioners in these fields. For these reasons, the empirical study of the public law enforcement system is less well developed than the empirical literature on the economic model of crime.

4.2. Policing strategies

The preceding section described the finding that increases in police reduce crime. Those studies investigated how the number of police employed affects crime, not how structuring the work of a given number of police affects crime. Like any labor allocation problem, additional police can presumably be assigned in relatively more or less productive ways. Empirical evidence on which policing strategies are the most effective at combating crime is to-date thin.

Criminologists have conducted numerous studies of how police perform their jobs and the efficacy of various policing strategies, such as community policing, "hot spot" policing, and crackdowns. Nagin (1998) and Sherman (2002) review this literature. Economists have devoted less energy to understanding which methods of policing are most effective, but a few exceptions exist. Grogger (2002), for example, studied the impact of gang ordinances in Los Angeles. He found that in neighborhoods subject to ordinances, violent crime fell by 5–10% in the year following their introduction, relative to neighborhoods without ordinances. Miles (2005) studied another law enforcement tool purported to increase the probability of apprehension, publicity. Using a sample of wanted fugitives, Miles found that publicity, as measured by an appearance on the television program "America's Most Wanted," correlated with a shorter time until apprehension.

Like social scientists in other disciplines, economists have documented and attempted to explain the large racial disparities in search and arrest rates. Donohue and Levitt (2001b) reported a relationship between the racial composition of police forces and the pattern of arrests. They found that increases in the number of minority police correlate with increases in the arrests of whites but not with the arrests of minorities. Similarly, increases in the number of white police correlate with increases in the arrests of minorities but not with the arrests of whites.

The fact that police stop and search African-Americans at much higher rates than whites has prompted allegations of "racial profiling," that police use race as a criteria in choosing whether to search. Knowles et al. (2001) developed a model to test whether racial prejudice or the goal of maximizing the number of successful searches explains the racial disparity in searches. Two economic theories of discrimination are relevant to racial profiling. Becker's (1957) model describes racial prejudice as a preference for searching members of a particular race. Under Arrow's (1973) model of "statistical discrimination," officers are not racially prejudiced and instead maximize the number of searches in which contraband is found. If race predicts criminality, police in the Arrowian equilibrium search African-Americans more often, even if police are not prejudiced. Perisco (2002) expanded this theoretical model of police searches to include a notion of fairness and the stigma of being searched.

Knowles et al. (2001) compared the rates at which searches yielded contraband across racial groups. If police are not racially prejudiced and merely maximize the number of successful searches, the rates at which searched persons are found carrying contraband should be equal across racial groups. Knowles et al. (2001) tested their hypothesis using a data set obtained from the ACLU containing all vehicle searches along an interstate highway in Maryland over a four-year period. The authors could not reject equality across racial groups in the rates at which police find contraband during searches. They also could not reject equality in the rates of successful searches across various other dimensions, such as gender, age of the vehicle, whether the vehicle is a luxury model, and whether the search occurred during the day or at night. The estimates of Knowles et al. (2001) suggest that racial disparities in vehicle searches result from a desire to maximize successful searches rather than from racial prejudice.

Several economists have criticized the Knowles et al. (2001) finding. Dharmapala and Ross (2003) argued that in a more generalized model including varying offenses levels and imperfect observability of offenders, the appropriate test requires stratifying the data by offense severity. Using the Knowles et al. (2001) data, Dharmapala and Ross (2003) found that African-Americans have higher guilty rates for felonies and whites have higher guilty rates for misdemeanors. Antonovics and Knight (2004) also argue that the ability of the Knowles et al. (2001) model to distinguish preference-based discrimination from statistical discrimination is not robust to generalization. Antonovics and Knight (2004) instead predict that if officers engage only in statistical discrimination, search decisions are independent of an officer's race (assuming police are randomly assigned geographically by race). Using data from the Boston Police Department, they showed that it is not. A consensus about the explanation for racial disparities in search rates has not emerged among economists, and the questions of race and policing are likely to remain an active and contentious area of research.

4.3. Prosecution of offenses

4.3.1. Criminal procedure

Criminal procedure, the legal rules that govern the prosecution, sentencing, and appeals of criminal defendants, is a large and complex field of law and legal scholarship. Economists have given the subject relatively little attention, and this omission is unfortunate. Economics is a natural mode of analysis for criminal procedure, because these rules set incentives for offenders, defendants, and other actors in the criminal justice system. In general terms, more exacting procedures make prosecutions more difficult. Criminal procedure affects the cost to the police and prosecutors of obtaining a conviction and the likelihood that an offender is sanctioned.

Recently, some researchers have attempted to study the empirical consequences of two major rules of criminal procedures: *Miranda v. Arizona*, 384 U.S. 436 (1966) and *Mapp v. Ohio*, 367 U.S. 643 (1961). These studies lack the econometric sophistication of recent work on the size of police forces and prison populations, but they represent a first forays of empirical economics into criminal procedure.

The Supreme Court's decision in *Miranda* is perhaps the most famous case in criminal procedure. *Miranda* held that interrogations of criminal defendants in the custody of police are presumptively compelled. This presumption can be dispelled only when the defendant receives the now famous four warnings: the right to remain silent, the consequence of not remaining silent, the right to counsel, and the right to court-appointed counsel at state expense. The remedy for a violation of *Miranda* is the suppression or exclusion at trial of the obtained statements. At the time of the decision, *Miranda* was enormously controversial, but in 2000, Chief Justice Rehnquist, wrote in *United States v. Dickerson*, 530 U.S. 428, 443 (2000), that the *Miranda* warnings have "become embedded in routine police practice to the point where the warnings have become part of the national culture." Despite its eventual acceptance, critics such as Justice White recognized that a potential cost of *Miranda* is a greater incidence of crimes because some potentially guilty offenders would escape conviction.[5] Although Justice White identified harms accruing from the loss of incapacitation, the increased likelihood that authorities cannot secure convictions attenuates deterrence as well.[6]

Following *Miranda*, numerous studies measured the decision's effect on police interrogations. Cassell (1996), Schulhofer (1996), and Thomas and Leo (2002) summarize this large literature. These studies observed police interrogations, conducted surveys and interviews about interrogations, and reviewed case files. Most studies, such as Leiken (1970) and Wald et al. (1967), concluded that police generally complied with the *Miranda* warning requirements. An unexpected finding was that after receiving the *Miranda* warnings, suspects typically waived their rights and communicated with interrogators. More recent studies suggest that these patterns have persisted in the thirty years since *Miranda*. Leo (1996, 1998), for example, reports that police regularly issue *Miranda* warnings but successfully induce 78% to 96% of suspects to waive their rights. However, suspects with criminal records are more likely to invoke their rights and terminate the interrogation (Leo, 1996; Cassell and Hayman, 1996).

Cassell (1996) attempted to calculate the number of convictions lost as a result of *Miranda*. After reviewing the literature, he asserted that *Miranda* caused a 16% reduction in the confession rate and 24% of convictions relied on confessions. The product of these figures suggests that *Miranda* caused a 3.8% reduction in convictions. Using the number of arrests as a proxy for the number of interrogations, Cassell claimed that *Miranda* resulted in the annual loss of 28,000 violent crime convictions and 79,000

[5] *Miranda*, 384 U.S. at 542–43 ("In some unknown number of cases the Court's rule will return a killer, a rapist or other criminal to the streets and to the environment which produced him, to repeat his crime whenever it pleases him. As a consequence, there will not be a gain, but a loss, in human dignity . . . There is, of course, a saving factor: the next victims are uncertain, unnamed, and unrepresented in this case") (White, J., dissenting).

[6] A popular misconception is that *Miranda* is intended to prevent the conviction of the innocent. Only one footnote in *Miranda* mentions the possibility that police may coerce innocent suspects into confessing. 384 U.S. at 455 n.24. If *Miranda* served to reduce erroneous convictions while raising the costs of enforcement, the effect on crime rates would be ambiguous. Greater accuracy in adjudication would enhance incentives for deterrence (Png, 1986), but higher adjudication costs would raise the raise the price of enforcement. However, the *Miranda* court repeatedly expressed concern that police manipulation may deprive suspects their "free choice." *See* 384 U.S. at 457, 458, 474, 465, 467, 478, 460.

property crime convictions. He also argued that *Miranda* weakened the terms of plea bargains in many other cases. Schulhofer (1996) disputed these figures. Placing the drop in confessions at 4.1% and the percentage of cases in which a confession was necessary for conviction at 19%, Schulhofer argued that *Miranda* resulted in the loss of only 0.78% of criminal cases. Moreover, Schulhofer thought these losses were likely only transitory, because police managed to adapt to *Miranda* and to elicit confessions despite its requirements.

Cassell and Fowles (1998) attempted a second measure of the cost of *Miranda*. Using an annual, national time-series for 1960–95, they regressed clearance rates (or the ratio of arrests to reported offenses) on a set of control variables, including an indicator variable for the years 1966–68. Cassell and Fowles claimed that the indicator variable captured the effect of *Miranda*, and their estimates showed that the clearance rates for robbery and property crimes declined during these years. Donohue (1998) and Feeney (2000) offer extensive criticisms of Cassell and Fowles's empirical methodology. Perhaps most obvious, a correlation in national time-series cannot support a causal inference because of possible omitted variables. Although interest in measuring the effect of *Miranda* remains strong, persuasive evidence remains elusive.

Another major Supreme Court decision on criminal procedure was *Mapp v. Ohio*, 367 U.S. 643 (1961). *Mapp* applied to the states the so-called "exclusionary rule," that evidence obtained by police in violation of the Fourth Amendment's prohibition against unreasonable searches and seizures, cannot be introduced, or must be excluded, at trial. Atkins and Rubin (2003) hypothesize that if *Mapp* raises the costs of police investigation, crime rises. Before *Mapp*, half of the states already had some form of exclusionary rule, and half permitted the admission at trial of evidence obtained in violation of the Fourth Amendment. This variation in the legal rule permitted Atkins and Rubin to compare the effect exclusionary rule on crime rates before and after *Mapp* as well as across states. In a panel of states covering the period 1948–67, their difference-in-differences estimates indicated that the exclusionary rule caused an increase in robbery and property crime rates by 4% to 8%. The estimates suggest that criminal procedure may be a significant influence on crime rates and that further research on the relationship between procedural safeguards and incentives for crime is needed.

4.3.2. Prosecutorial behavior

After arrest, the prosecutor makes the determination of whether the defendant should face punishment. Landes (1971, 1974) modeled prosecutors as maximizing the sum of expected sentences in their case loads. Some legal scholars, such as Stuntz (2004), dispute that sentence-maximization is prosecutors' objective and instead emphasize the wide degree of discretion prosecutors retain in choosing whom to prosecute and which offenses to charge.

The assumptions about prosecutorial preferences are especially relevant to the debates over the desirability of "mandatory minimums," or statutes that set compulsory lower limits on sentences. Under Landes's conception of prosecutorial behavior, mandatory minimums should raise the length of the average sentence, but under the alternative framework, prosecutors may wield their discretion to undo the statutory requirements. Tonry (1996), for example, claims that prosecutors perceive mandatory sentences as unduly harsh and prosecutors avoid them by charging offenses that are similar but do not trigger application of the mandatory sentence. Empirical economic studies of prosecutorial behavior, while few in number, suggest that prosecutors exercise considerable discretion, and their evidence on whether prosecutors maximize expected sentence lengths is mixed.

Kessler and Piehl (1998) studied the effect of California's Proposition 8, a law passed in 1982 that required mandatory minimum sentences for certain violent offenses. They compared sentencing outcomes before and after the passage of the law among three groups of defendants: those convicted of crimes subject to the mandatory minimum; those convicted of crimes factually similar but not subject to the mandatory minimum; and those convicted of crimes factually different and not subject to the mandatory minimum. They found, for example, that the average sentence for robbery, an offense subject to the mandatory minimum, rose by 50% following the law's passage. The average sentence for grand larceny, an offense that is factually similar to robbery but not subject to robbery's mandatory minimum, rose slightly. The average sentence for drug possession, a factually different offense that is not subject to the mandatory minimum, fell slightly. Kessler and Piehl interpreted their results as evidence that prosecutorial discretion is not as unfettered as often alleged and that when prosecutors exercise their discretion, they do so to increase rather than decrease sentences.

Bjerk (2003) studied sentencing outcomes in several states following the passage of three-strikes laws. His difference-in-difference estimates implied that when a defendant was arrested on a felony charge that would trigger application of the third strike enhancement, prosecutors were roughly twice as likely to charge a misdemeanor instead. The fraction of cases in which prosecutors exercised this discretion remained small, however. It did not exceed 10%. Despite the prosecutorial tendency to reduce charged offense, the three-strike law appeared to raise mean sentence of defendants with qualifying criminal histories and current convictions.

Other key aspects of the criminal justice system have received scant attention from empirical economists. For example, especially since the Supreme Court's decisions in *United States v. Lopez*, 514 U.S. (1995), and *United States v. Morrison*, 529 U.S. 598 (2000), criminal law scholars have been particularly interested in the trend of "federalizing" offenses that previously only state laws proscribed. In contrast, only one empirical economic study has considered the differences between the state and federal prosecutions. Glaeser et al. (2000) compared the characteristics of drug crime defendants in two systems. The authors found that federal defendants are on average older and more highly educated and that they are more likely to be female, white, Hispanic, and married. Federal defendants were also more likely to have experience in managerial or technical employment and less likely to have a prior conviction. They are more often charged with drug distribution rather than possession and hire private attorneys with greater frequency. Further empirical study of the prosecutorial behavior and of the differences between the state and federal systems is clearly needed.

4.3.3. Bail system

Bail is an important mechanism for securing the appearance of defendants at trial, and economists have examined several aspects of it. The first economic studies of the bail system focused on whether the likelihood of a defendant's flight declined with larger bail amounts. In a sample of criminal defendants in New York City, Landes (1973) found that offense type and bail amounts correlated with the probability of jumping bail. Clarke et al. (1976) reached a somewhat different conclusion by examining defendants in Charlotte, NC. They claimed that the strictness of supervision during release and court delay were strongly related to the likelihood of a defendant failing to appear but that offense type and demographic characteristics bore no systematic relationship to flight risk. Myers (1981) showed that the existence of a plea bargain greatly affected the estimates. In a sample of New York defendants, Myers found that after controlling for plea bargains, the probability of flight was only weakly correlated with court delay but strongly and negatively correlated with larger bail amounts. Helland and Tabarrok (2004) demonstrate that bounty hunters are much more effective than public law enforcement in reducing bail jumping and catching those who skip bail.

Ayres and Waldfogel (1994) studied criminal defendants in New Haven, CT, and found that bail amounts for African-American and Hispanic men were substantially higher than for whites and women, even after controlling for observable characteristics. Their estimates also indicated that African-Americans and Hispanics paid significantly lower interest rates on their bail bonds. Ayres and Waldfogel argue that because bond dealers willingly accept lower interest rates from minorities for use of their capital, the possibility that minorities pose greater flight risks cannot justify their higher bail amounts. Ayres and Waldfogel believe their estimates are initial evidence of racial prejudice in setting bail.

4.4. Sentencing

4.4.1. Effect of attempts to cabin judicial discretion

The central issue in sentencing policy over the past fifteen years has been the efficacy and desirability of the U.S. Sentencing Guidelines. Before the introduction of the Guidelines in 1987, sentencing judges had effectively unfettered discretion to sentence defendants to terms falling anywhere within the statutory range, which often encompassed not only the length of sentence, but also the choice between probation and prison. The advantage of so-called "indeterminate sentencing" is that it affords judges the opportunity to tailor sentences to individual defendants and to extend leniency where appropriate. However, critics of the system alleged that broad judicial discretion compromised the principle of equal treatment under law by producing dissimilar treatment of similar defendants. Under this system, the sentences of defendants convicted of the same offense and having similar criminal histories exhibited wide variation. In addition, these variations were often correlated with race.[7]

In response to these problems, the Sentencing Guidelines attempted to restrain judicial discretion in sentencing while retaining sufficient flexibility to permit individual sentences in each case (Breyer, 1988). The Guidelines narrowed the range of permissible sentences and allowed deviations from those ranges only for a certain specified reasons. Under the Guidelines, a grid largely determines sentences (United States Sentencing Commission, 2003). The vertical axis measures the severity of the defendant's criminal history, and the horizontal axis measures the severity of the current offense. A complex set of rules governs calculation of the criminal history category and offense level. Each box on the grid contains a sentence range. A judge may "depart" from the Guidelines, or impose a sentence other than the range specified on the grid, only in an atypical case that presents factors or circumstances that the Guidelines do not consider. The purpose of the Guidelines is thus to reduce the influence of the identity of the sentencing judge or the demographic characteristics of the defendant in the determination of the sentence.

Anderson et al. (1999) studied whether the Guidelines were successful in reducing inter-judge variation in sentences by comparing sentences before and after implementation of the Guidelines. The authors estimated that the expected difference in sentence between any two judges fell by roughly five months after the implementation of the Guidelines. However, they also found that over this time period, the average sentence rose from 24 to 35 months. LaCasse and Payne (1999) obtained roughly similar results by comparing cases in two federal courts in New York before and after the introduction of the Guidelines. They reported that the amount of variation in sentences attributable to the identity of the judge did not change for plea bargains and even rose for cases that proceeded to trial. Because both these studies used samples that included only federal defendants, all of whom become subject to the Guidelines, their comparisons were not made relative to a control group. Thus, whether their estimates capture the effect of the Guidelines or other contemporaneous changes in sentencing policy, such as the introduction of mandatory minimums for certain drug offenses, is uncertain.

A particularly relevant dimension of judicial discretion pertains to race. Mustard (2001) and Bushway and Piehl (2001) review the large literature on the role of the

[7] Courts have been famously unmoved by statistical evidence of racial disparities in sentencing outcomes and have perceived the issue as a responsibility of the legislature. For example, in *McCleskey v. Kemp*, 481 U.S. 297, 320 (1987), the Supreme Court declined to intervene in case in which a defendant presented a statistical study showing that, even after controlling for observable differences, the defendants in Georgia accused of murdering whites were roughly four times more likely to receive a death sentence than were defendants accused of murdering blacks. The Court concluded, "McCleskey's arguments are best presented to the legislative bodies."

defendant's race and other demographic characteristics in sentencing outcomes. Sociologists and criminologists have studied this question extensively, but economists are more recent entrants to this literature. Mustard (2001) and Bushway and Piehl (2001) attempted to determine whether demographic characteristics remain significant factors in the post-Guidelines period. Both reported that they are. Mustard found that, among federal defendants, males, African-Americans, and persons with less than high school education received longer sentences. The sentences of drug-trafficking defendants exhibited the widest variation. The disparities resulted principally from departures from the Guidelines rather than within-range differences. Bushway and Piehl studied sentences in Maryland, a state with sentencing guidelines analogous to the federal system. They found that African-Americans receive sentences 20% longer after controlling for other factors. Bushway and Piehl found that departures from the guidelines were most substantial in the portion of the sentencing grid that recommended longer sentences and that African-American defendants were concentrated in this portion of the grid. Although the studies do not provide a persuasive measure of the whether the Guidelines have strengthened or weakened importance of demographic factors, they suggest that the Guideline's success in cabining the influence of those factors is at best qualified.

4.4.2. *Empirical tests of optimal type of penalty*

Economists have also tested whether the manner in which society enforces its laws is efficient or socially optimal. An optimal penalty is the combination of a probability of apprehension and a set of punishments that minimize the costs of maintaining a given level of expected penalty. Although the early theoretical literature emphasized the cost difference between policing and fines (Becker, 1968; Polinsky and Shavell, 1979) and between fines and incarceration (Polinsky and Shavell, 1984), empiricists have examined the imposition of different kinds of sanctions. Some studies consider the sentences meted out to individual criminal defendants. Waldfogel (1995) predicted that efficiency would require the maximum use of fines before the use of incarceration, because fines are costless and imprisonment is costly. In a sample of federal fraud defendants, he found that ability to pay correlated positively with fines and negatively with the length of prison terms. Waldfogel and Meade (1998) reported similar evidence of the optimizing tendency in judges' choice between fines and imprisonment. The authors also tested whether the Sentencing Guidelines raise the costs of punishment by limiting a judge's ability to substitute fines for prison. They estimate that the Guidelines raise the cost of punishment by roughly 5% of the total cost of imprisoning federal inmates.

Economists have also noted that the total punishment for offending consists of more than just the state-imposed sentence. Arrest and conviction may damage a defendant's reputation and impair his opportunities in the legitimate sector. Some economists tested this prediction by comparing offenders' earnings before and after their convictions. Lott (1990, 1992b) found that convicts suffered declines in post-release earnings consistent with collateral consequences of conviction, such as the loss of professional licenses and damage to reputation.

Most empirical research on reputational sanctions investigates the consequences of criminal activity by corporations rather than individuals, because the price of a corporation's stock provides a ready measure of its reputation. Karpoff and Lott (1993) reported that news of an investigation or an indictment of a corporation for fraud resulted in large declines in market value. In contrast, the stock price effect for a regulatory violation was small and statistically insignificant. Their estimates show that, relative to the market sanction, government-imposed criminal penalties for fraud are very small. Karpoff et al. (1999) also found evidence of reputational sanctions in a sample of defense contractors. The authors showed that the stock of a defense contractor experienced negative abnormal returns following press reports of a fraud investigation, indictment, or suspension. Their results indicated that the market penalties are not uniform for all firms, because peripheral defense contractors had the most negative abnormal returns. The share price declines may reflect a reluctance to conduct business with an offending corporation. Alexander (1999) found that allegations of criminal activity by a corporation correlate with reports of suspended or severed customer relationships and management or employee turnover. These reports are more likely when the offense harms customers, as in fraud, than when it injures third parties, as in environmental crimes.

Just as economists studied the effect of the Sentencing Guidelines on individual defendants, economists have also examined their impact on the sentencing of corporations. The Guidelines for corporate defendants, like those for individuals, attempted to constrain judicial discretion in determining sentences, but unlike the Guidelines for individuals, they also sought to increase the level of sanctions. Parker and Atkins (1999) compared corporate criminal fines before and after the Guidelines, and after controlling for the magnitude of the harm, they found no statistically significant change in corporate monetary penalties. In contrast, Alexander et al. (1999a) claim that total penalties for corporate crime were higher following the Guidelines. Fines, which are governed by the Guidelines, are roughly five times higher, but even non-fine penalties, which are not constrained by the Guidelines, rose following the implementation of the Guidelines. Differences in the empirical specification and the data used may explain the conflicting results of these two studies. Alexander et al. (1999b) note that Parker and Atkins's use of restitution as a proxy for harm may mask any effect on fines, because the Guidelines sought to increase restitution as well as fines. Alexander et al., unlike Parker and Atkins, construct their own data set of corporate sentences, because variable definitions in the Sentencing Commission's corporate sentencing data changed over time. Generally, the lack of available data continues to hamper empirical studies of corporate crime and limits economists' understanding of how corporate offending differs from individual criminal behavior. Alexander et al. (2000) discuss the shortcomings of even the Commission's post-Guidelines data.

4.5. Incarceration strategies

Social scientists, including economists, have produced surprisingly little evidence on the effectiveness of incarceration in reducing recidivism. Economists have not yet examined whether participants in prison education programs, boot camps, or faith-based programs, for example, have lower rates of re-offending than the general prison population. An obstacle in any such study would be the self-selection of participants into these programs. Even if this econometric difficulty were overcome, the consensus that the criminal justice system does not advance rehabilitation and Donohue and Siegelman's (1998) pessimistic view of the intervention programs bode ill for the efficacy of prison programs.

Even if prison programs successfully rehabilitated offenders, they might be in tension with the goals of deterrence. Katz et al. (2003) analyzed the relationship between crime rates and prison conditions. They find that more severe prison conditions, as measured by the death rate among prisoners, correlate with declines in crime rates. Whether these penal goals conflict and whether prison programs are effective awaits future research.

4.6. Post-release

In addition to studying whether the criminal justice system deters crime, some economists have examined the system's impact on those who were not deterred. The effect of contact with the criminal justice system on subsequent behavior has gained renewed relevance as many of the prisoners incarcerated in the 1980's and 1990's approach their release dates. Numerous studies, such as Langan and Levine (2002), report high recidivism and re-incarceration rates, and Raphael and Stoll (2003) examine the relationship between crime rates and the size of the parolee population. Using a repeated cross-section of states between 1977 and 2001, the authors find that larger parole populations correlate with higher rates of all major felonies. However, they estimate that the increases in crime caused by the release of the marginal parolee are lower than the reductions accruing from the incarceration of the marginal inmate. Their results do not distinguish whether the marginal parolee has a lower criminal propensity than the marginal inmate, or whether the conditions of parole themselves reduce criminal activity.

Economists have looked most often for an impact of the criminal justice system on offenders' labor market outcomes. The effect may vary with the degree of the contact with the criminal justice system: arrest, conviction, incarceration, and length of any incarceration. The prevailing wisdom is that these contacts have a negative effect on employment and earnings. Arrest and conviction might adversely affect an individual's labor market prospects because involvement with the criminal justice system signals the individual's untrustworthiness, and human capital may deteriorate during incarceration. However, economists have found no evidence of a deleterious effect on human capital and only limited support for the stigma hypothesis among certain white-collar offenders.

Grogger (1995) studied the effect of arrests on a sample of young men in California and found that the impact was small and short-lived. According to his estimates, any effect of arrest dissipated after 18 months. In a sample of young British offenders, Nagin and Waldfogel (1995) found no effect of self-reports of criminal activity on labor market outcomes, but a conviction correlated, curiously, with an increase in an offender's legitimate sector earnings. They hypothesized that conviction relegates offenders to spot-market employment that raise earnings in the short term. Nagin and Waldfogel (1998) tested this further hypothesis by examining the relationship of a first-time conviction and the earnings of federal offenders. Their estimates indicate that a first conviction has a positive effect on earnings for offenders under age 25 but an increasingly negative impact over age 30.

Other studies test the effect of a prison sentence and its length, rather than the effect of a conviction. Results of these studies generally vary by the offense of conviction. Waldfogel (1994) reported that federal defendants sentenced to prison for fraud or larceny experienced significant reductions in employment rates and income relative to similarly convicted defendants who did not receive prison terms. Lott (1992b) also found that the length of a prison sentence correlates with post-release earnings declines for federal defendants convicted of fraud and embezzlement. However, Lott (1992a) reported no significant relationship between post-release earnings and sentence length for federal drug offenders. Kling (2002) obtained similar results in a sample of California convicts. He found that drug and violent offenders have low average earnings, but post-release their earnings increase and do not vary by the length of time served. In contrast, white collar offenders who were incarcerated earned 10–30% less five to eight years after release than similarly convicted defendants who did not serve time. Economists have thus demonstrated that the effect of contact with the criminal justice system on labor market outcomes is more complex than previously thought and that its effect varies over time and by demographic group.

5. Conclusion

Empirical economists have made substantial contributions to the understanding of criminal behavior. In particular, they have provided evidence supporting the deterrence model, and they have shown that incapacitation is at work as well. Economists have begun to assess the relative importance of deterrence and incapacitation, but more research is needed on this question. The evidence of the crime-reducing effect of police and prisons exhibits consistency across different methodological approaches. In contrast, estimates of the deterrent effect of other penalties, such as capital punishment and three-strikes laws, are less robust and suggest that claims of large effects from these policies may be spurious.

Empirical economists have made less progress in studying the criminal justice system itself. These studies lack the econometric sophistication of deterrence research, perhaps because the type of data that would permit such tests remain unavailable. These areas remain important avenues for future empirical research on criminal punishment.

References

Alexander, C.R. (1999). “On the Nature of the Reputational Penalty for Corporate Crime: Evidence”. Journal of Law and Economics 42, 489–526.

Alexander, C.R. (2000). "Evaluating Trends in Corporate Sentencing: How Reliable Are the U.S. Sentencing Commission's Data?". Federal Sentencing Reporter 13, 108–113.

Alexander, C.R., Arlen, J., Cohen, M.A. (1999a). "Regulating Corporate Criminal Sanctions: Federal Guidelines and the Sentencing of Public Firms". Journal of Law and Economics 42 (pt. 2), 271–300.

Alexander, C.R., Arlen, J., Cohen, M.A. (1999b). "The Effect of Federal Sentencing Guidelines on Penalties for Public Corporations". Federal Sentencing Reporter 12, 20–35.

Anderson, J.M., Kling, J.R., Stith, K. (1999). "Measuring Inter-Judge Sentencing Disparity Before and After the Federal Sentencing Guidelines". Journal of Law and Economics 42 (1 pt. 2), 271–307.

Antonovics, K.L., Knight, B.G. (2004). "A New Look at Racial Profiling: Evidence from the Boston Police Department", National Bureau of Economic Research Working Paper, No. 10634, July.

Arrow, K.J. (1973). "The Theory of Discrimination". In: Ashenfelter, O., Rees, A. (Eds.), Discrimination in Labor Markets. Princeton University Press, Princeton, pp. 3–33.

Atkins, R.A., Rubin, P.H. (2003). "Effects of Criminal Procedure on Crime Rates: Mapping Out the Consequences of the Exclusionary Rule". Journal of Law and Economics 46, 157–180.

Ayres, I., Donohue III, J.J. (2003). "Shooting Down the 'More Guns, Less Crime' Hypothesis". Stanford Law Review 55, 1193–1312.

Ayres, I., Levitt, S.D. (1998). "Measuring Positive Externalities from Unobservable Victim Precautions: An Empirical Analysis of Lojack". Quarterly Journal of Economics 113 (1), 43–77.

Ayres, I., Waldfogel, J. (1994). "A Market Test for Race Discrimination in Bail Setting". Stanford Law Review 46, 987–1046.

Bailey, W.C. (1982). "Capital Punishment and Lethal Assaults on Police". Criminology 19, 608–625.

Bar-Gill, O., Harel, A. (2001). "Crime Rates and Expected Sanctions: The Economics of Deterrence Revisited". Journal of Legal Studies 30, 485–501.

Bebchuk, L.A., Kaplow, L. (1992). "Optimal Sanctions When Individuals Are Imperfectly Informed About the Probability of Apprehension". Journal of Legal Studies 21, 365–370.

Beccaria, C. (1992). An Essay on Crimes and Punishments. International Pocket Library, Brookline Village, MA (originally published 1770).

Beck, A. (1989). "Recidicivism of Prisoners Released in 1983", Bureau of Justice Statistics Special Report. Bureau of Justice Statistics, Washington, DC.

Becker, G.S. (1957). The Economics of Discrimination. University of Chicago Press, Chicago.

Becker, G.S. (1968). "Crime and Punishment: An Economic Approach". Journal of Political Economy 76, 169–217.

Bentham, J. (1973). "An Introduction to the Principles of Morals and Legislation". In: The Utilitarians. Anchor Books, Garden City, NJ, pp. 5–398 (originally published 1789).

Bjerk, D.J. (2003). "Making the Crime Fit the Punishment: The Role of Prosecutorial Discretion Under Mandatory Minimum Sentencing", McMaster University Mimeo.

Blumstein, A., Cohen, J., Roth, J.A., Visher, C.A. (1986). Career Criminals and Criminal Careers. National Academy Press, Washington, DC.

Bowles, R., Garoupa, N. (1997). "Casual Police Corruption and the Economics of Crime". International Review of Law and Economics 17, 75–87.

Bushway, S., Piehl, A.M. (2001). "Judging Judicial Discretion: Legal Factors and Racial Discrimination in Sentencing". Law & Society Review 35, 733–764.

Breyer, S. (1988). "The Federal Sentencing Guidelines and Key Compromises on Which They Rest". Hoftsra Law Review 17, 1–51.

Cameron, S. (1988). "The Economics of Deterrence: A Survey of Theory and Evidence". Kyklos 41, 301–323.

Cameron, S. (1994). "A Review of Econometric Evidence on the Effects of Capital Punishment". Journal of Socio-Economics 23 (1), 197–214.

Cassell, P.G. (1996). "*Miranda's* Social Costs: An Empirical Reassessment". Northwestern University Law Review 90, 387–499.

Cassell, P.G., Fowles, R. (1998). "Handcuffing the Cops? A Thirty-Year Perspective on *Miranda's* Harmful Effects on Law Enforcement". Stanford Law Review 50, 1055–1145.

Cassell, P.G., Hayman, B.G. (1996). “Police Interrogation in the 1990s: An Empirical Study of the Effects of *Miranda*”. U.C.L.A. Law Review 43, 839–931.

Chaiken, J., Chaiken, M. (1982). Varieties of Criminal Behavior. Rand, Santa Monica, Calif.

Clarke, S.H., Freeman, J.L., Koch, G.G. (1976). “Bail Risk: A Multivariate Analysis”. Journal of Legal Studies 10, 341–385.

Clotfelter, C. (1978). “Private Security and the Public Safety”. Journal of Urban Economics 5, 388–402.

Cloward, R., Ohlin, L. (1960). Delinquency and Opportunity: A Theory of Delinquent Gangs. Free Press, New York.

Cohen, A. (1955). Delinquent Boys. Free Press, New York.

Cook, P.J., Moore, M.H., Braga, A.A. (2002). “Gun Control”. In: Wilson, J.Q., Petersilia, J. (Eds.), Crime. Institute for Contemporary Studies, San Francisco, pp. 291–330.

Corman, H., Mocan, H.N. (2000). “A Time-Series Analysis of Crime and Drug Use in New York City”. American Economic Review 90, 584–604.

Dezhbakhsh, H., Rubin, P.H., Shepherd, J.M. (2003). “Does Capital Punishment Have a Deterrent Effect? New Evidence from Postmoratorium Panel Data”. American Law and Economics Review 5 (2), 344–376.

Dharmapala, D., Ross, S.L. (2003). “Racial Bias in Motor Vehicle Searches: Additional Theory and Evidence”, University of Connecticut Department of Economics Working Paper, No. 2003-12R, December.

Di Tella, R., Schargrodsky, E. (2004). “Do Police Reduce Crime? Estimates using the Allocation of Police Forces after a Terrorist Attack”. American Economic Review 94, 115–133.

Donohue III, J.J. (1998). “Did *Miranda* Diminish Police Effectiveness?”. Stanford Law Review 50, 1147–1180.

Donohue III, J.J., Levitt, S.D. (2001a). “The Impact of Legalized Abortion on Crime”. Quarterly Journal of Economics 116, 379–420.

Donohue III, J.J., Levitt, S.D. (2001b). “The Impact of Race on Policing and Arrests”. Journal of Law and Economics 44 (pt. 1), 367–394.

Donohue III, J.J., Siegelman, P. (1998). “Allocating Resources among Prisons and Social Programs in the Battle Against Crime”. Journal of Legal Studies 27, 1–43.

Donohue III, J.J., Wolfers, J. (2006). “Uses and Abuses of Empirical Evidence in the Death Penalty Debate”. Stanford Law Review 58, 791–846.

Duggan, M. (2000). “More Guns, More Crime”. Journal of Political Economy 109 (5), 1086–1114.

Ehrlich, I. (1973). “Participation in Illegitimate Activities: A Theoretical and Empirical Investigation”. Journal of Political Economy 81, 531–567.

Ehrlich, I. (1975). “The Deterrent Effect of Capital Punishment: A Question of Life and Death”. American Economic Review 65, 397–417.

Ehrlich, I. (1977). “Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence”. Journal of Political Economy 85, 741–788.

Ehrlich, I. (1981). “On the Usefulness of Controlling Individuals: An Economic Analysis of Rehabilitation, Incapacitation, and Deterrence”. American Economic Review 71, 307–332.

Ehrlich, I., Liu, Z. (1999). “Sensitivity Analyses of the Deterrence Hypothesis: Let’s Keep the Econ in Econometrics”. Journal of Law and Economics 42 (part 2), 455–487.

Evans, W.N., Owens, E. (2004). “Flypaper COPS”, University of Maryland Working Paper, September.

Fagan, J., West, V., Holland, J. (2003). “Reciprocal Effects of Crime and Incarceration in New York City Neighborhoods”. Fordham Urban Law Journal 30, 1551–1602.

Federal Bureau of Investigation (2002). Crime in the United States. Government Printing Office, Washington, DC.

Feeney, F. (2000). “Police Clearance: A Poor Way to Measure the Impact of *Miranda* on Police”. Rutgers Law Review 32, 1–114.

Fisher, F., Nagin, D. (1978). “On the Feasibility of Identifying the Crime Function in a Simultaneous Equations Model of Crime and Sanctions”. In: Blumstein, A., Nagin, D., Cohen, J. (Eds.), Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates. National Academy of Sciences, Washington, DC, pp. 361–399.

Garoupa, N. (1999). "Optimal Law Enforcement with Dissemination of Information". European Journal of Law and Economics 7, 183–196.

Glaeser, E.L., Kessler, D.P., Piehl, A.M. (2000). "What Do Prosecutors Maximize? An Analysis of the Federalization of Drug Crimes". American Law and Economics Review 2 (2), 259–290.

Glaeser, E.L., Sacerdote, B. (1999). "Why Is There More Crime in Cities?". Journal of Political Economy 107 (6 pt. 2), 225–258.

Glaeser, E.L., Sacerdote, B., Scheinkman, J.A. (1996). "Crime and Social Interactions". Quarterly Journal of Economics 111 (2), 507–548.

Glassner, B., Ksander, M., Berg, B., Johnson, B. (1983). "A Note on the Deterrent Effect of Juveniles vs. Adult Jurisdiction". Social Problems 31, 219–221.

Granger, C. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". Econometrica 37, 424–438.

Greenwood, P., Rydell, C.P., Abrahamse, A., Caulkins, J., Chiesa, J., Model, K., Klein, S. (1994). Three Strikes and You're Out: Estimated Benefits and Costs of California's New Mandatory Sentencing Law. RAND, Santa Monica, CA.

Grogger, J. (1995). "The Effect of Arrests on the Employment and Earnings of Young Men". Quarterly Journal of Economics 110, 51–72.

Grogger, J. (2002). "The Effects of Civil Gang Injunctions on Reported Violent Crime: Evidence from Los Angeles County". Journal of Law and Economics 45 (1), 69–90.

Heckman, J.J. (1995). "Lessons from the Bell Curve". Journal of Political Economy 103 (5), 1091–1120.

Helland, E., Tabarrok, A. (2004). "The Fugitive: Evidence on Public versus Private Enforcement from Bail Jumping.". Journal of Law and Economics 47 (1), 93–122.

Hernnstein, R., Murray, C. (1994). The Bell Curve. Free Press, New York.

Hirschi, T. (1969). Causes of Delinquency. University of California Press, Berkeley.

Hirschi, T., Hindelang, M.J. (1977). "Biological and Psychophysiological Factors in Criminality". In: Glaser, D. (Ed.), Handbook of Criminology. Rand McNally, Chicago, pp. 101–173.

Jolls, C.M., Sunstein, C.R., Thaler, R. (1998). "A Behavioral Approach to Law and Economics". Stanford Law Review 50, 1471–1550.

Kaplow, L. (1990). "Optimal Deterrence, Uninformed Individuals, and Acquiring Information About Whether Acts Are Subject to Sanctions". Journal of Law, Economics, and Organization 6, 93–128.

Kaplow, L., Shavell, S. (2002). "Economic Analysis of Law". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 3. Elsevier, New York, pp. 1661–1784.

Karpoff, J.M., Lee, D.S., Vendrzyk, V.P. (1999). "Defense Procurement Fraus, Penalties, and Contractor Influence". Journal of Political Economy 107, 809–842.

Karpoff, J.M., Lott, J.R. Jr. (1993). "The Reputational Penalty Firms Bear from Committing Criminal Fraud". Journal of Law and Economics 36, 757–802.

Katz, L., Levitt, S.D., Shustorovich, E. (2003). "Prison Conditions, Capital Punishment, and Deterrence". American Law and Economics Review 5, 318–343.

Kelling, G., Pate, T., Dieckman, D., Brown, C. (1974). The Kansas City Preventative Patrol Experiment: A Summary Report, 1974. Police Foundation, Washington, DC.

Kennedy, P. (1998). A Guide to Econometrics. MIT Press, Cambridge, MA.

Kennedy, D.M., Braga, A.A., Piehl, A.M. (1996). "Youth Violence in Boston: Serious Youth Offenders and a Use-Reduction Strategy". Law and Contemporary Problems 59, 474–483.

Kessler, D.P., Levitt, S.D. (1999). "Using Sentence Enhancements to Distinguish between Deterrence and Incapacitation". Journal of Law and Economics 17 (1), 343–363.

Kessler, D.P., Piehl, A.M. (1998). "The Role of Discretion in the Criminal Justice System". Journal of Law, Economics, and Organization 14 (2), 256–276.

Klein, L., Forst, B., Filatov, V. (1978). "The Deterrent Effect of Capital Punishment: An Assessment of the Estimates". In: Blumstein, A., Nagin, D., Cohen, J. (Eds.), Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates. National Academy of Sciences, Washington, DC, pp. 336–360.

Kling, J.R. (2002). "The Effect of Prison Sentence Length on the Subsequent Employment and Earnings of Criminal Defendants", Princeton University Working Paper.

Knowles, J., Persico, N., Todd, P. (2001). "Racial Bias in Motor Vehicle Searches: Theory and Evidence". Journal of Political Economy 109 (1), 203–232.

Kuziemko, I., Levitt, S.D. (2004). " An Empirical Analysis of Imprisoning Drug Offenders". Journal of Public Economics 88 (9–10), 2043–2066.

LaCasse, C., Payne, A.A. (1999). "Federal Sentencing Guidelines and Mandatory Minimum Sentences: Do Defendants Bargain in the Shadow of the Judge?". Journal of Law and Economics 42, 245–269.

Landes, W.M. (1971). "An Economic Analysis of Courts". Journal of Law and Economics 14, 61–107.

Landes, W.M. (1973). "The Bail System: An Economic Approach". Journal of Legal Studies 2, 79–105.

Landes, W.M. (1974). "Legality and Reality: Some Evidence on Criminal Procedure". Journal of Legal Studies 3, 287–337.

Langan, P.A., Levine, D.J. (2002). Recidivism of Prisoners Released in 1984. Bureau of Justice Statistics, Washington, DC.

Larson, R.C. (1976). "What Happened to Patrol Operations in Kansas City?". Evaluation 3, 123–177.

Leamer, E. (1983). "Let's Take the Con out of Econometrics". American Economic Review 73, 31–43.

Leiken, L.S. (1970). "Police Interrogation in Colorado: The Implementation of *Miranda*". Denver Law Journal 47, 1–53.

Leo, R.A. (1996). "Inside the Interrogation Room". Journal of Criminal Law and Criminology 86, 266–303.

Leo, R.A. (1998). "*Miranda* and the Problem of False Confessions". In: Leo, R.A., Thomas III, G.C. (Eds.), The *Miranda* Debate: Law, Justice, and Policing. Northeastern University Press, Boston, pp. 271–282.

Levitt, S.D. (1996). "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Litigation". Quarterly Journal of Economics 111, 319–325.

Levitt, S.D. (1997a). "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime". American Economic Review 87 (3), 270–290.

Levitt, S.D. (1997b). "Private Information as an Explanation for the Use of Jail Sentences Instead of Fines". International Review of Law and Economics 17, 179–192.

Levitt, S.D. (1998a). "Why Do Increased Arrest Rates Appear to Reduce Crime: Deterrence, Incapacitation, or Measurement Error?". Economic Inquiry 36, 353–372.

Levitt, S.D. (1998b). "Juvenile Crime and Punishment". Journal of Political Economy 106, 1156–1185.

Levitt, S.D. (2002). "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Reply". American Economic Review 92, 1244–1250.

Levitt, S.D. (2004). "Understanding Why Crime Fell in the 1990's: Four Factors that Explain the Decline and Six that Do Not". Journal of Economic Perspectives 18 (1), 163–190.

Levitt, S.D., Venkatesh, S. (2000). "An Economic Analysis of a Drug-Selling Gang's Finances". Quarterly Journal of Economics 115, 755–789.

Lombroso, C. (1876). L'uomo Delinquete (The Criminal Man). Putnam, New York, reprint 1911.

Lott, J.R. Jr. (1990). "The Effect of Conviction on the Legitimate Income of Criminals". Economics Letters 34, 381–385.

Lott, J.R. Jr. (1992a). "Do We Punish High Income Criminals Too Heavily?". Economic Inquiry 30, 583–608.

Lott, J.R. Jr. (1992b). "An Attempt at Measuring the Total Monetary Penalty From Drug Convictions: The Importance of an Individual's Reputation". Journal of Legal Studies 21, 159–187.

Lott, J.R. (1998). More Guns, Less Crime. University of Chicago Press, Chicago.

Lott, J.R. Jr., Landes, W. (1999). "Multiple Victim Public Shootings, Bombings, and Right-to-Carry Concealed handgun Laws: Constrasting Private and Public Law Enforcement", University of Chicago Law School, John M. Olin Law & Economics Working Paper No. 73.

Lott, J.R. Jr., Mustard, D.B. (1997). "Crime, Deterrence, and Right-to-Carry Concealed Handguns". Journal of Legal Studies 26 (1), 1–68.

Ludwig, J. (1998). "Concealed-Gun-Carrying Laws and Violent Crime: Evidence from State Panel Data". International Review of Law and Economics 18, 239–254.

Marvell, T., Moody, C. (1994). "Prison Population Growth and Crime Reduction". Journal of Quantitative Criminology 10, 109–140.

Marvell, T., Moody, C. (1996). “Specification Problems, Police Levels, and Crime Rates”. Criminology 34, 609–646.

McCormick, R., Tollison, R. (1984). “Crime on the Court”. Journal of Political Economy 92, 223–235.

McCrary, J. (2002). “Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Comment”. American Economic Review 92 (4), 1236–1243.

McManus, W. (1985). “Estimates of the Deterrent Effect of Capital Punishment: The Importance of the Researcher’s Prior Beliefs”. Journal of Political Economy 93, 417–425.

Merton, R.K. (1968). Social Theory and Social Structure. Free Press, New York. originally published in 1957.

Miles, T. (2005). “Estimating the Effect of ‘America’s Most Wanted’: A Duration Analysis of Fugitives”. Journal of Law & Economics 48, 281–306.

Mocan, H.N., Gittings, R.K. (2003). “Getting Off Death Row: Commuted Sentences and the Deterrent Effect of Capital Punishment”. Journal of Law and Economics 46, 453–478.

Mustard, D.B. (2001). “Racial, Ethnic, and Gender Disparities in Sentencing: Evidence from the U.S. Federal Courts”. Journal of Law and Economics 44, 85–314.

Myers, S.L. Jr. (1981). “The Economics of Bail Jumping”. Journal of Legal Studies 10, 381–392.

Nagin, D. (1998). “Criminal Deterrence Research: A Review of Evidence and a Research Agenda for the Outset of the 21st Century”. In: Tonry, M. (Ed.), Crime and Justice, vol. 23, pp. 1–42.

Nagin, D., Waldfogel, J. (1995). “The Effects of Criminality and Conviction on the Labor Market Status of Young British Offenders”. International Review of Law and Economics 15, 109–126.

Nagin, D., Waldfogel, J. (1998). “The Effect of Conviction on Income Through the Life Cycle”. International Review of Law and Economics 18, 25–40.

O’Brien, R. (1985). Crime and Victimization Data. Sage, Beverly Hills, CA.

Parker, J.S., Atkins, R.A. (1999). “Did the Corporate Criminal Sentencing Guidelines Matter? Some Preliminary Empirical Observations”. Journal of Law and Economics 42, 423–453.

Passell, P., Taylor, J. (1977). “The Deterrent Effect of Capital Punishment: Another View”. American Economic Review 67, 445–451.

Perisco, N. (2002). “Racial Profiling, Fairness, and Effectiveness of Policing”. American Economic Review 92 (5), 1472–1497.

Piehl, A.M., DiIulio, J.J. Jr. (1995). “‘Does Prison Pay?’ Revisited: Returning to the Crime Scene”. The Brookings Review xx (Winter), 21–25.

Polinsky, A.M., Rubinfeld, D. (1991). “A Model of Optimal Fines for Repeat Offenders”. Journal of Political Economy 24, 89–99.

Polinsky, A.M., Shavell, S. (1979). “The Optimal Tradeoff Between the Probability and Magnitude of Fines”. American Economic Review 68, 880–891.

Polinsky, A.M., Shavell, S. (1984). “The Optimal Use of Fines and Imprisonment”. Journal of Public Economics 24, 89–99.

Polinsky, A.M., Shavell, S. (2001). “Corruption and Optimal Law Enforcement”. Journal of Public Economics 81, 1–24.

Polinsky, A.M. Shavell, S. (2007). This volume.

Png, I.P.L. (1986). “Optimal Subsidies and Damages in the Presence of Judicial Error”. International Review of Law and Economics 6, 101–105.

Raine, A. (2002). “The Biological Basis of Crime”. In: Wilson, J.Q., Petersilia, J. (Eds.), Crime. Institute for Contemporary Studies, San Francisco, pp. 43–74.

Raphael, S., Stoll, M. (2003). “Parolees and Crime”, paper prepared for the Brookings-Wharton Conference on Urban Affairs.

Rasmusen, McAdams (2007). This volume.

Reckless, W. (1961). The Crime Problem. Appleton-Century-Crofts, New York.

Sah, R.K. (1991). “Social Osmosis and Patterns of Crime”. Journal of Political Economy 99 (6), 1272–1295.

Saltzman, L., Paternoster, R., Waldo, G., Chiricos, T. (1982). “Deterrent and Experiential Effects: The Problem of Causal Order in Perceptual Deterrence Research”. Journal of Research in Crime and Delinquency 19, 172–189.

Schulhofer, S.J. (1996). "*Miranda's* Practical Effect: Substantial Benefits and Vanishingly Small Social Costs". Northwestern University Law Review 90, 500–563.

Shah, S.A., Roth, L.H. (1974). "Biological and Psychophysiological Factors in Criminality". In: Glaser, D. (Ed.), Handbook of Criminology. Rand McNally, Chicago, pp. 101–173.

Shavell, S. (1987). "A Model of Optimal Incapacitation". American Economic Review 77, 107–110.

Shavell, S. (1991). "Individual Precautions to Prevent Theft: Private versus Socially Optimal Behavior". International Review of Law and Economics 11, 123–132.

Sheldon, W. (1940). The Varieties of Human Physique. Harper and Row, New York.

Shepherd, J.M. (2002). "Police, Prosecutors, Criminals, and Determinate Sentencing: The Truth About Truth-in-Sentencing Laws". Journal of Law and Economics 45 (2), 505–530.

Shepherd, J.M. (2004). "Murders of Passion, Execution Delays, and the Deterrence of Capital Punishment". Journal of Legal Studies 33, 283–322.

Sherman, L.W. (2002). "Fair and Effective Policing". In: Wilson, J.Q., Petersilia, J. (Eds.), Crime. Institute for Contemporary Studies, San Francisco, CA, pp. 383–412.

Stuntz, W.J. (2004). "Plea Bargaining and Criminal Law's Disappearing Shadow". Harvard Law Review 117, 2548–2569.

Sutherland, E. (1956). The Professional Thief. Phoenix Press, Chicago, originally published by University of Chicago Press, 1937.

Thomas, G.C., Leo, R.A. (2002). "The Effects of *Miranda v. Arizona*: 'Embedded' in Our National Culture?". In: Tonry, M. (Ed.), Research in Crime and Justice. University of Chicago Press, Chicago, pp. 203–271.

Tonry, M. (1996). Sentencing Matters. Oxford University Press, New York.

United States Sentencing Commission (2003). Guidelines Manual. Government Printing Office, Washington, DC.

Wald, M. (1967). "Interrogations in New Haven: The Impact of *Miranda*". Yale Law Journal 76, 1519–1648.

Walfogel, J. (1994). "The Effect of Criminal Conviction on Income and Trust Reposed in the Workmen". Journal of Human Resources 29, 62–81.

Waldfogel, J. (1995). "Are Fines and Prison Terms Used Efficiently? Evidence on Federal Fraud Offenders". Journal of Law and Economics 38 (1), 107–139.

Waldfogel, J., Meade, J. (1998) "Do Sentencing Guidelines Raise the Cost of Punishment?", National Bureau of Economic Research Working Paper, No. 6361, January.

Zimring, F., Hawkins, G. (1991). The Scale of Imprisonment. University of Chicago Press, Chicago.

This page intentionally left blank

PART II:

ADDITIONAL AREAS OF THE LEGAL SYSTEM

This page intentionally left blank

Chapter 8

ENVIRONMENTAL LAW

RICHARD L. REVESZ[*]

School of Law, New York University

ROBERT N. STAVINS[†]

John F. Kennedy School of Government, Harvard University, and Resources for the Future

Contents

[*] Revesz is Dean of the New York University School of Law, where he is the Lawrence King Professor of Law, and Director, Program on Environmental Regulation, Center on Environmental and Land Use Law.

[†] Stavins is the Albert Pratt Professor of Business and Government at the John F. Kennedy School of Government, Director of the Harvard Environmental Economics Program, and a University Fellow, Resources for the Future.

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)01008-0

Abstract

This chapter provides an economic perspective of environmental law and policy. We examine the ends of environmental policy, that is, the setting of goals and targets, beginning with normative issues, notably the Kaldor–Hicks criterion and the related method of assessment known as benefit–cost analysis. We examine this analytical method in detail, including its theoretical foundations and empirical methods of estimation of compliance costs and environmental benefits. We review critiques of benefit–cost analysis, and examine alternative approaches to analyzing the goals of environmental policies.

We examine the means of environmental policy, that is, the choice of specific policy instruments, beginning with an examination of potential criteria for assessing alternative instruments, with particular focus on cost-effectiveness. The theoretical foundations and experiential highlights of individual instruments are reviewed, including conventional, command-and-control mechanisms, market-based instruments, and liability rules. Three cross-cutting issues receive attention: uncertainty; technological change; and distributional considerations. We identify normative lessons in regard to design, implementation, and the identification of new applications, and we examine positive issues: the historical dominance of command-and-control; the prevalence in new proposals of tradeable permits allocated without charge; and the relatively recent increase in attention given to market-based instruments.

We also examine the question of how environmental responsibility is and should be allocated among the various levels of government. We provide a positive review of the responsibilities of Federal, state, and local levels of government in the environmental realm, plus a normative assessment of this allocation of regulatory responsibility. We focus on three arguments that have been made for Federal environmental regulation: competition among political jurisdictions and the race to the bottom; transboundary environmental problems; and public choice and systematic bias.

Keywords

environmental economics, environmental law, efficiency, cost-effectiveness, benefit–cost analysis, environmental federalism

JEL classification: K320, Q280, Q380, Q480

1. Introduction

An economic perspective can provide clarity regarding the causes and consequences of environmental degradation, and thereby provide insights regarding public policies intended to protect the environment. This is true both with regard to normative and positive assessments of environmental policies. Despite this value, an economic perspective is by no means a perfect substitute for other legitimate perspectives on environmental law and policy, whether from the natural sciences, from ethics, or from other disciplines. Rather, an economic perspective is a valuable complement to such views. Indeed, over the past several decades, as the attention given to environmental issues in the United States has grown, greater consideration has also been given to the efficiency, cost-effectiveness, and distributional equity of laws and regulations intended to protect the environment.[1]

In an effort to be rigorous in our review while keeping the treatment to reasonable length, we have imposed limits on the scope of our coverage. First, we focus on pollution control, and do not consider natural resource management, despite the fact that these two areas are closely related. Second, we concentrate our attention on environmental protection efforts at the federal level in the United States, and do not examine state, local, or international regulatory efforts.

We begin with the core question of whether and why environmental regulation is needed, considering the fact that under many conditions unconstrained markets produce socially desirable outcomes. What about in the environmental sphere? Under what specific circumstances will governmental intervention be appropriate? The fundamental theoretical argument for government activity in the environmental realm is that pollution is a classic example of an externality—an unintended consequence of market decisions, which affects individuals other than the decision maker. Because firm-level decisions do not take into account full social costs, pollutant emissions tend to be higher than socially efficient levels. As environmental quality is thus naturally under-provided by competitive markets, a possible role arises for government regulation. The traditional theoretical solution to the externality problem was long thought to be to force private actors to "internalize" the full costs of their actions. The primary advocate of this view was Arthur Pigou, who in *The Economics of Welfare* (1920) proposed that the government should impose a tax on emissions equal to the cost of the related damages at the efficient level of control.

A critical response to the Pigovian perspective was provided by Ronald Coase in his seminal article, *The Problem of Social Cost* (1960). Coase made three key points.

[1] We follow the standard definition of an *efficient* environmental policy as being one that involves a target—such as a 50 percent reduction in sulfur dioxide (SO_2) emissions—that maximizes the difference between social benefits and social costs, i.e., a target level at which marginal benefits and marginal costs are equated. By *cost-effective* policies, we refer to those that take (possibly inefficient) targets as given by the political process, but achieve those targets with policy instruments—such as a tradeable permit system in the SO_2 case—that minimize aggregate costs. Assessments of the *distributional* implications of environmental policies include analyses of the incidence of costs and of benefits.

First, he argued that even if A's actions inflict harm on B, it is not the case that A's actions should necessarily be restrained, because A's harm of B is really "a problem of a reciprocal nature" which arises because of the simultaneous presence of two parties. For example, the problem of a factory that emits fumes that harm a nearby laundry is not caused solely by the factory. Protecting the laundry by enjoining the fumes would impose harm on the factory, just as protecting the factory by not enjoining its actions would impose harm on the laundry.

Second, Coase demonstrated that in a bargaining environment without transaction costs, parties will reach socially desirable agreements; and third, that the overall amount of pollution will be independent of the legal rules (assignment of property rights) chosen to structure their relationship. For example, if the legal regime enjoined pollution, but the harm to the factory were greater than the harm that the laundry would have suffered in the absence of such an injunction, the parties will enter into a contract under which, in return for a payment, the laundry will agree not to exercise its right to seek an injunction. Conversely, if the legal regime allows the pollution but the resulting harm to the laundry is greater than the harm that the injunction would impose on the factory, the parties will enter into a contract under which, again in return for a payment, the factory would agree not to pollute. Thus, regardless of the initial legal rule, bargaining will produce two results: (1) it will lead to the same amount of pollution; and (2) it will lead to the maximization of social welfare. Of course, the choice of legal rules can determine which party makes payments and which party receives them, a distributional concern, though not one of efficiency.

These three points are jointly characterized as the Coase Theorem. The Theorem may be said to hold if there are no transaction costs,[2] no wealth or income effects,[3] private rather than public goods, and no third-party impacts (i.e., all affected parties participate in the negotiation). At least some of these conditions are unlikely to hold in the case of most environmental problems. Hence, private negotiation will not—in general—fully internalize environmental externalities. And when market transactions—including Coasian bargaining—do not generate socially efficient allocations of resources, government regulation may be necessary to improve environmental quality.

On the other hand, although government regulation may be necessary to improve environmental quality when market transactions fail to generate socially efficient allocations of resources, such regulation is by no means sufficient to improve welfare or even environmental quality. This is because government regulation itself may not be efficient, that is, government may under-regulate or over-regulate, and/or it may regulate in ways that require unnecessarily large costs of compliance.[4]

[2] Coase recognized that transaction costs can be significant, and could prevent efficient negotiated outcomes. When transaction costs are great, the choice of legal rule will affect the amount of pollution and hence the level of social welfare.

[3] That is, when the size of the payments is sufficiently small relative to the firms' or individuals' incomes or wealth that payment and receipt has no effect on respective supply and demand functions.

[4] The public choice literature in economics suggests specific reasons for "government failure," analogous to market failure. See our application of the economic theory of politics in section 3.2, for example.

We continue in section 2 of this chapter with an examination of the ends of environmental policy, that is, the setting of goals and targets. We begin with an examination of normative issues, notably the Kaldor–Hicks criterion and the related method of benefit–cost analysis. We examine this analytical method in detail, including its theoretical foundations and various approaches to the estimation of compliance costs and environmental benefits. We include a review of critiques of benefit–cost analysis, and examine alternative approaches to analyzing the goals of environmental policies. The section closes with a positive survey of the efforts of the Federal governmental to use these analytical methods.

In section 3, we turn to the means of environmental policy, that is, the choice of specific policy instruments. We begin with normative issues, and examine potential criteria for assessing alternative instruments, with particular focus on cost-effectiveness. The theoretical foundations and experiential highlights of individual instruments are reviewed, beginning with conventional, command-and-control, and then turning to economic incentive or market-based instruments. In the latter category, we consider pollution charges, tradeable permit systems, market friction reductions, and government subsidy reductions. We also consider the role of liability rules. Three cross-cutting issues merit particular attention: implications of uncertainty for instrument choice; effects of instrument choice on technological change; and distributional considerations. From this review, we identify a set of normative lessons in regard to design, implementation, and the identification of new applications. The section closes with an examination of positive issues, including three phenomena we seek to explain: the historical dominance of command-and-control; the prevalence in new proposals of tradeable permits allocated without charge; and the relatively recent increase in attention given to market-based instruments.

In section 4, we turn to the question of how environmental responsibility is and should be allocated among the levels of government. We offer a positive review of the responsibilities of Federal, state, and local levels of government in the environmental realm, plus a normative assessment of this allocation of regulatory responsibility. We examine three arguments that have been made for federal environmental regulation: competition among political jurisdictions and the "race to the bottom;" transboundary environmental problems; and public choice problems. In section 5, we conclude.

2. Setting goals and targets: the ends of environmental policy

If it is deemed appropriate for government to become involved in environmental protection, how intensive should that activity be? In real-world environmental policy, this question becomes how stringent should our environmental goals and targets be? For example, should we cut back sulfur dioxide (SO_2) emissions by 10 million tons, or would a 12 million ton reduction be better? In general, how clean is clean enough? How safe is safe enough?

2.1. Normative issues and analysis

Most economists would argue that economic efficiency—achieved when the difference between benefits and costs is maximized—ought to be one of the fundamental criteria for evaluating environmental protection efforts. Because society has limited resources to spend, benefit–cost analysis can help illuminate the trade-offs involved in making different kinds of social investments. In practice, there are significant challenges, in large part because of inherent difficulties in measuring benefits and costs. In addition, concerns about fairness and process merit consideration, because public policies inevitably involve winners and losers, even when aggregate benefits exceed aggregate costs.

2.1.1. Criteria for environmental policy evaluation

More than 100 years ago, Vilfredo Pareto enunciated the well-known normative criterion for judging whether a social change—possibly induced by public policy—makes the world better off: a change is *Pareto efficient* if at least one person is made better off, and no one is made worse off (1896). This criterion has considerable normative appeal, but virtually no public policies meet the test of being true Pareto improvements, since there are inevitably some in society who are made worse off by any conceivable change. Nearly fifty years later, Nicholas Kaldor (1939) and John Hicks (1939) postulated a more pragmatic criterion that seeks to identify "potential Pareto improvements:" a change is defined as welfare-improving if those who gain from the change could—in principle—fully compensate the losers, with (at least) one gainer still being better off.

The Kaldor–Hicks criterion—a test of whether total social benefits exceed total social costs—is the theoretical foundation for the use of the analytical device known as benefit–cost (or net present value) analysis. Neither the Pareto efficiency criterion nor the Kaldor–Hicks criterion calls for support for *any* policy for which benefits are greater than costs. Rather, the key is to identify the policy for which the positive difference between benefits and costs is greatest; otherwise it would be possible to identify another policy that would represent a further (potential) Pareto improvement.

If the objective is to maximize the difference between benefits and costs (net benefits), then the related level of environmental protection (pollution abatement) is defined as the efficient level of protection:

$$\max_{\{q_i\}} \sum_{i=1}^{N} [B_i(q_i) - C_i(q_i)] \rightarrow q_i^* \tag{1}$$

where q_i is abatement by source i ($i = 1$ to N), $B_i(\cdot)$ is the benefit function for source i, $C_i(\cdot)$ is the cost function for the source, and q_i^* is the efficient level of protection (pollution abatement). The key necessary condition that emerges from the maximization problem of equation (1) is that marginal benefits be equated with marginal costs.

The Kaldor–Hicks criterion is clearly more practical than the strict Pareto criterion,[5] but its normative standing is less solid and has been attacked from various quarters. Although basic economic (utility) theory posits that individual well-being is a function of the satisfaction of individual preferences, this notion has been debated in other disciplines, including psychology and philosophy.[6] In addition, questions have been raised about whether social gains and losses can be expressed through the simple aggregation of welfare changes of individuals. Some have argued that other factors should be considered in a measure of social well-being, and that criteria such as distributional equity should trump efficiency considerations in some collective decisions (Kelman, 1981a; Sagoff, 1993).[7] Many economists do not disagree with this assertion, and indeed have noted that the Kaldor–Hicks criterion should be considered neither a necessary nor a sufficient condition for public policy (Arrow et al., 1996b).

At the heart of the claim that the Kaldor–Hicks criterion lacks normative standing for public decision making is the lack of any guarantee that compensation can or will be paid. Gains and losses to individuals can be aggregated in a variety of ways, but the standard method of aggregation is a simple sum, an approach that can be problematic if there is diminishing marginal utility of income or individual utility is dependent on the overall societal distribution of income. Under such conditions, policies can pass a potential Pareto improvement test, but decrease overall societal well-being, or vice-versa. Thus, some of the debate may be understood as focusing on the compatibility of the efficiency and distributional equity criteria.[8] The general view from economics is that other criteria in addition to efficiency can and should be employed by policy makers, but that the existence of such criteria does not invalidate the efficiency criterion, which should remain part of social decision-making (Arrow et al., 1996b; Kopp, Krupnick, and Toman, 1997).[9]

Many proposed and implemented environmental policies involve real trade-offs between equity and efficiency, and both international and national policy bodies have demonstrated concern for ensuring that groups such as low-income citizens, ethnic minorities, and future generations do not bear “disproportionate” shares of the costs of

[5] If a proposal fails the (weaker) Kaldor–Hicks test, it cannot pass the Pareto test. Hence, at a minimum, the Kaldor–Hicks criterion can be used to weed out the worst policies, that is, those that cannot make the world better off in the Pareto efficiency sense.

[6] See, for example, Scanlon (1991) for a philosophical critique, and Kopp (1993) for a response. More broadly, see Adler and Posner (2001).

[7] For a contrasting view, see Kaplow and Shavell (2001, 2002a), who argue that any policy assessment that accords importance to non-utility criteria violates the Pareto principle and, thus, is subject to powerful criticism. See the discussion in section 3.1.3.3, below, of distributional considerations.

[8] See: Polinsky (1971, 1980).

[9] Data limitations sometimes reduce the reliability of economic benefit estimates, thus reducing the efficacy of benefit–cost analysis and the operational content of the efficiency criterion. Economics can still aid in decision making through the cost-effectiveness criterion, where an environmental target is taken as given, and the least-cost means of achieving that target are identified. We consider cost-effectiveness analysis later, in the context of normative analysis of policy instrument choice.

environmentally related actions.[10] While it is conceivable to combine the goals of equity and efficiency using a social welfare function to arrive at a single metric (Bergson, 1938; Jorgenson, 1997), the information constraints and collective choice caveats have been acknowledged (Arrow, 1963, 1977; Sen, 1970). The consensus, at least within the realm of environmental policy, is that efficiency and equity ought to be evaluated separately (U.S. Environmental Protection Agency, 2000a), but there is no consensus on specific criteria that might be used to rank alternatives from an equity perspective.[11]

In recent years, there has been much debate among economists, and between economists and nearly everyone else regarding the meaning of the frequently employed concept of "sustainability." Ecologists and many others outside the economics profession have taken sustainability to be the unique and comprehensive criterion that can and should guide global development. In contrast, economists have tended to define sustainability as being purely about intertemporal distribution, that is, intergenerational equity.[12] As such, most economists have viewed sustainability as no more than one element of a desirable development path.

A broader notion of sustainability, with considerable appeal outside of economics, combines two components—dynamic efficiency and intergenerational equity (Stavins, Wagner, and Wagner, 2003). Thus, a sustainable path is one that is both efficient and non-decreasing in utility over time. Much as a potential Pareto-improvement in the Kaldor–Hicks sense can yield Pareto optimality when combined with appropriate compensation of losers by winners, so too can dynamic efficiency lead to the ambitious goal of sustainability when combined with appropriate intergenerational transfers. The implication is that much as practical economic analyses often resort to seeking potential Pareto-improvements (see the following section), so too might intertemporal economic analyses focus on dynamic efficiency, leading to the possibility, at least, of sustainability.

[10] Executive Order 12898 (1994) provides a mandate for Federal agencies to make "environmental justice" part of their missions by considering possible negative effects of proposed policies on minority and low-income populations (U.S. Council on Environmental Quality, 1997). In the international realm, as early as 1987, the Brundtland Commission defined development as sustainable "when it meets the needs of the present without compromising the ability of future generations to meet theirs" (World Commission on Environment and Development, 1987).

[11] In the intertemporal realm, Rawls' (1971) cognitive device of a "veil of ignorance" is insightful, but not operational. Farrow (1998) proposed a modified benefit–cost test for intergenerational equity that emphasized actual compensation rather than potential improvement.

[12] For example, Arrow et al. (2004) make a clear distinction between "optimality," defined as the maximized discounted present value of future well being, and sustainability, defined as "the maintenance or improvement of well being over time." One exception is provided by Asheim, Buchholz, and Tungodden (2001), who impose so-called efficiency and equity axioms and show that if social preferences fulfill these two axioms, any optimal path will lead to an efficient and non-decreasing path, thus implicitly including dynamic efficiency in the definition of sustainability. For a broader discussion of sustainability and optimality, see Pezzey (1992) and Weitzman (2003), and for a review of the major issues involved, see Pezzey and Toman (2001, 2002).

2.1.2. Benefit–cost analysis of environmental regulations

While conceptually straightforward, the soundness of empirical benefit–cost analysis rests upon the availability of reliable estimates of social benefits and costs,[13] including estimates of the social discount rate.

2.1.2.1. Discounting Decisions made today typically have impacts both now and in the future. In the environmental realm, many of the future impacts are from policy-induced improvements, and so in this context, future benefits (as well as costs) of policies are discounted (Goulder and Stavins, 2002). The present value of net benefits (PVNB) is defined as:

$$\mathrm{PVNB} = \sum_{t=0}^{T} \{(B_t - C_t) \cdot (1 + r)^{-t}\} \tag{2}$$

where B_t are benefits at time t, C_t are costs at time t, r is the discount rate, and T is the terminal year of the analysis. A positive PVNB means that the policy or project has the potential to yield a Pareto improvement (meets the Kaldor–Hicks criterion).[14] Thus, carrying out benefit–cost or "net present value" (NPV) analysis requires discounting to translate future impacts into equivalent values that can be compared. In essence, the Kaldor–Hicks criterion provides the rationale both for benefit–cost analysis and for discounting.

Choosing the discount rate to be employed in an analysis can be difficult, particularly where impacts are spread across a large number of years involving more than a single generation.[15] The rate chosen can have a significant effect if there are large differences among policies in the timing of benefits and costs. In general, benefits and costs should be discounted at the social discount rate—the relative valuation placed by society on future consumption presently sacrificed. In theory, the social discount rate could be derived by aggregating the individual time preference rates of all parties affected by a policy. Under idealized conditions, the market interest rate would reflect the

[13] Early volumes on benefit–cost analysis include those by Mishan (1976) and Stokey and Zeckhauser (1978); and a recent text is by Boardman et al. (2001). One of the earliest applications to environmental and natural resource policy was by Eckstein (1958).

[14] Neither benefit/cost ratios (dividing benefits by costs) nor internal rates of return (the interest rate that results in the present value of benefits being equal to the present value of costs) provide satisfactory alternatives to the net present value criterion, because—among other reasons—neither takes into account scale, and hence both can fail to make proper comparisons among policies using the Kaldor–Hicks criterion. Benefit–cost ratios have the additional problem that the ranking of projects is sensitive to the fundamentally arbitrary judgment of whether an environmental externality is considered to be an increment to costs or a decrement to benefits.

[15] Useful surveys include Lind (1982) and Portney and Weyant (1999). An important distinction is whether a publically-mandated policy or project calls for public or private spending. On the effects of this distinction on the choice of discount rate, see, for example: Scheraga and Sussman (1998).

marginal rate of time preference of individuals, but the presence of taxes, risk, liquidity constraints, limited information, and other imperfections means that the social discount rate is not reflected by any particular market rate (Newell and Pizer, 2004).

Alternatives to constant exponential discounting have received consideration. Evidence from market behavior and from experimental economics indicates that individuals may employ lower discount rates for impacts of larger magnitude, higher discount rates for gains than for losses, and rates that decline with the time span being considered (Cropper, Aydede, and Portney, 1994; Cropper and Laibson, 1999). In particular, there has been both empirical and theoretical support for the use of hyperbolic discounting and similar approaches with declining discount rates over time (Ainslie, 1991; Weitzman, 1994, 1998), but most of these approaches suffer from the problem that they would imply inconsistent decisions over time. Declining discount rates based on uncertainty in future rates, however, need not suffer from the time-inconsistency problem (Newell and Pizer, 2003a).

The choice of discount rate can be particularly important in the case of environmental problems with very long time horizons, such as global climate change, radioactive waste disposal, groundwater pollution, and biodiversity preservation (Revesz, 1999). Choosing an intergenerational rate is difficult, because the preferences of future generations are unknowable, and ethical questions arise about trading off the well-being of future generations. Approaches to intergenerational discounting have been considered in two conceptual categories. One relies on a social planner approach, which seeks to maximize the utilities of present and future generations, based on a social welfare function (Lind, 1995; Schelling, 1995; Arrow et al., 1996a). A second approach is based on the preferences of existing individuals, and assumes that one of the allocation decisions these individuals must make is about the welfare of future generations (Shefrin and Thaler, 1988; Cropper, Aydede, and Portney, 1992; Rothenberg, 1993; Schelling, 1995).

What discount rates are actually employed by government agencies? The general answer is a "large range." For many years, the U.S. Office of Management and Budget (OMB) required the use of a 7 percent real rate for Regulatory Impact Analyses (RIAs), despite the fact that this seems high compared with advice from economists regarding the social discount rate, which would place it in the range of 2 to 3 percent. Why did this persist? One possible rationale was that OMB believed that agencies want their policies, programs, and projects to go forward, and so will tend to exaggerate benefits relative to costs, and that OMB tried to counteract this effect by using a higher discount rate.[16] In any event, reforms put in place by OMB in September of 2003 included the use of a 3 percent real discount rate for intragenerational analyses and lower, unspecified rates for intergenerational contexts (U.S. Office of Management and Budget, 2003).

Several general principles are worth noting. First, it is generally appropriate to employ the same discount rate for benefits and costs. Second, if private capital investments

[16] This rationale assumes that the policies in question have the time profile of typical investments, that is, up-front costs and delayed benefits.

will be displaced by public projects, this should be taken into account in estimates of future benefits and costs prior to discounting. Third, estimates of future benefits and costs that may be uncertain or involve risk should be adjusted accordingly (such as through the use of certainty-equivalents), but the discount rate itself should not be changed to account for risk or uncertainty. Fourth, sensitivity analysis using alternative discount rates should be carried out.

2.1.2.2. Benefit concepts and taxonomies If an environmental change matters to any person—now or in the future—then it should, in principle, show up in an economic assessment.[17] Thus, the economic concept of environmental benefits is considerably broader than most non-economists would think.[18] The environment can be viewed as a form of natural asset that provides service flows used by people in the production of goods and services, such as agricultural output, human health, recreation, and more amorphous goods such as quality of life. This effect is analogous to the manner in which real physical capital assets provide service flows used in manufacturing. As with real physical capital, a deterioration in the natural environment (as a productive asset) reduces the flow of services the environment is capable of providing.

Protecting the environment usually involves active employment of capital, labor, and other scarce resources. Using these resources to protect the environment means they are not available to be used for other purposes. Hence, the economic concept of the value or benefit of environmental goods and services is couched in terms of society's willingness to make trade-offs between competing uses of limited resources, and in terms of aggregating over individuals' willingness to make these trade-offs. Thus, the benefits of an environmental policy are defined as the collection of individuals' willingness to pay (WTP) for the reduction or prevention of environmental damages or individuals' willingness to accept (WTA) compensation to tolerate such environmental damages. In theory, which measure of value is appropriate for assessing a particular policy depends upon the related assignment of property rights, the nature of the status quo, and whether the change being measured is a gain or a loss, but under a variety of conditions, the difference between the two measures may be expected to be relatively small (Willig, 1976).[19] Empirical evidence suggests larger than expected differences between willingness to pay and willingness to accept (Cummings, Brookshire, and Schulze, 1986;

[17] This reflects the anthropocentric view employed in economics, which does not include welfare incurred by other living creatures, unless it (indirectly) affects humans. For a recent argument involving non-anthropocentric values, see Ariansen (1998).

[18] For a summary of myths that non-economists seem to have regarding economics in the environmental realm, and a set of responses thereto, see: Fullerton and Stavins (1998).

[19] The difference depends on the magnitude of the impact, as well as the related income elasticity of demand. Consumers' surplus, derived from the observed Marshallian demand curve, provides a close approximation for equivalent and compensating variations (Willig, 1976). Willig's analysis is of price changes, but Randall and Stoll (1980) showed that similar results hold for quantity changes. For reviews of empirical studies of willingness-to-pay and willingness-to-accept measures, see: Horowitz and McConnell (2002, 2003). For examinations of the relationship between Willig's conditions and weak complementarity, see: Bockstael and McConnell (1993); Palmquist (2005); and Smith and Banzhaf (2004).

Fisher, McClelland, and Schulze, 1988). Theoretical explanations include psychological aversion to loss and poor substitutes for environmental amenities. In particular, Hanneman (1991) demonstrated that for quantity changes, the less perfect the substitutes that are available for a public good, the greater the expected disparity between WTP and WTA.

The benefits people derive from environmental protection are numerous and diverse. From a biophysical perspective, such benefits can be categorized as being related to human health (mortality and morbidity), ecological impacts (both market and non-market),[20] or materials damage. From an economic perspective, a critical distinction is between use value and non-use value. In addition to the direct benefits (use value) people receive through protection of their health or through use of a natural resource, people also derive passive or non-use value from environmental quality, particularly in the ecological domain. For example, an individual may value a change in an environmental good because she wants to preserve the option of consuming it in the future (option value) or because she desires to preserve the good for her heirs (bequest value). Still other people may envision no current or future use by themselves or their heirs, but still wish to protect the good because they believe it should be protected or because they derive satisfaction from simply knowing it exists (existence value).[21]

2.1.2.3. Cost concepts and taxonomies In the environment context, the economist's notion of cost, or more precisely, opportunity cost, is a measure of the value of whatever must be sacrificed to prevent or reduce the risk of an environmental impact. Hence, the costs of environmental policies are the forgone social benefits due to employing scarce resources for environmental policy purposes, instead of putting those resources to their next best use.[22]

A taxonomy of environmental costs can be developed, beginning with the most obvious and moving towards the least direct (Jaffe et al., 1995). First, many policy makers and much of the general public identify the on-budget costs to government of administering (monitoring and enforcing) environmental laws and regulations as the cost of environmental regulation. This meets the notion of opportunity cost, since administering environmental rules involves the employment of resources (labor and capital) that could

[20] It is important to distinguish between ecosystem functions (for example, photosynthesis) and the environmental services produced by ecosystems that are valued by humans (Freeman, 1997). The range of these services is great, including obvious environmental products such as food and fiber, and services such as flood protection, but also including the quality of recreational experiences, the aesthetics of the landscape, and such desires (for whatever reasons) as the protection of marine mammals.

[21] Option value and existence value should not be thought of as being additive, since option value is defined from a framework that holds expected utility constant; this is not the case with existence (and bequest) value. The contemporary concept of non-use value relates to what was previously most often characterized as existence value. See: Graham (1981); Bishop (1982); and Smith (1987).

[22] Costs and benefits are thus two sides of the same coin. The cost of an environmental-protection measure may be defined as the gross decrease in benefits (consumer and producer surpluses) associated with the measure and with any price and/or income changes that may result (Cropper and Oates, 1992).

otherwise be used elsewhere. But economic analysts also include as costs the capital and operating expenditures associated with regulatory compliance. Indeed, these typically represent a substantial portion of the overall costs of regulation, although a considerable share of compliance costs for some regulations falls on governments rather than private firms.[23] Additional direct costs include legal and other transaction costs, the effects of refocused management attention, and the possibility of disrupted production.

Next, there are what have sometimes been called "negative costs" of environmental regulation, including the beneficial productivity impacts of a cleaner environment and the potential innovation-stimulating effects of regulation.[24] General equilibrium or multi-market effects associated with discouraged investment[25] and retarded innovation constitute another important layer of costs, as do the transition costs of real-world economies responding over time to regulatory changes.[26]

2.1.2.4. Cost estimation methods The merits of alternative empirical cost estimation methods are related to the magnitude of the various categories of costs outlined above. Methods of direct compliance cost estimation, which measure the costs to firms of purchasing and maintaining pollution-abatement equipment plus costs to government of administering a policy, are acceptable when behavioral responses, transitional costs, and indirect costs are small. Partial and general equilibrium analysis allow for the incorporation of behavioral responses to changes in public policy. Partial equilibrium analysis of compliance costs incorporates behavioral responses by modeling supply and/or demand in major affected markets, but assumes that the effects of a regulation are confined to one or a few markets. This may be satisfactory if the markets affected by the policy are small in relation to the overall economy, but if an environmental policy is expected to have large consequences for the economy, general equilibrium analysis is required.

General equilibrium cost estimation methods include both input-output models and computable general equilibrium models. Input-output analysis quantifies the flow of goods and services in an economy using fixed-coefficient relationships (Leontief, 1966, 1970), and is limited in its usefulness by restrictive assumptions of constant returns to scale, fixed prices, and fixed producer and consumer behavior. Computable general equilibrium (CGE) models relax these assumptions at the cost of greater data

[23] One example in the United States is the federal regulation of contaminants in drinking water, the cost of which is borne primarily by municipal governments.

[24] The notion that environmental regulation can foster economic growth is a controversial one among economists. For a debate on this proposition, see: Porter and van der Linde (1995); and Palmer, Oates, and Portney (1995).

[25] For example, if a firm chooses to close a plant because of a new regulation (rather than installing expensive control equipment), this would be counted as zero cost in narrow compliance-cost estimates, but it is obviously a real cost.

[26] If a policy will result in only small changes in consumer and producer behavior, real resource and regulatory costs will represent the bulk of costs. But when behavioral responses are expected to be sizeable, social welfare costs associated with losses in consumer and producer surplus due to a rise in prices or a decrease in output can be significant.

requirements.[27] The potential importance of tax-interaction effects (Goulder, Perry, and Burtraw, 1997), described below in section 3.1.3.3, highlight the value of employing CGE models for comprehensive cost analysis.

How well have cost estimation methods performed in practice? In a retrospective examination of 28 environmental and occupational safety regulations, Harrington, Morgenstern, and Nelson (2000) found that fourteen had produced *ex ante* cost estimates that exceeded actual *ex post* costs.[28] But these errors were mainly due to overestimates of the quantity of emissions reduction that would occur. In terms of per-unit abatement costs, overestimation and underestimation were equally common, although for regulations that used economic incentives, per-unit costs were consistently overestimated. Harrington, Morgenstern, and Nelson (2000) attributed this to technological innovation which was stimulated by these market-based instruments, and which thereby reduced abatement costs.[29]

2.1.2.5. Benefit estimation methods Empirical methods of economic valuation were originally developed in the context of changes in individuals' incomes and in prices faced in the market. Over the past thirty years, these methods have been extended to accommodate changes in the qualities of goods, to public goods that are shared by individuals, and to other non-market services such as environmental quality and human health.[30] With markets, consumers' decisions about how much of a good to purchase at different prices reveal useful information regarding the surplus consumers gain. With non-market environmental goods, it is necessary to infer this willingness to trade off other goods (or monetary amounts) for environmental services using other methods. A repertoire of techniques has been developed in two broad categories: revealed preference (indirect measurement) and stated preference (direct questioning).

2.1.2.5.1. Revealed preference methods Whenever possible, it is preferable to measure trade-offs by observing actual decisions made by consumers in real markets. In limited situations, researchers can observe relationships that exist between a non-marketed, environmental good and some good that has a market price. In this case, individuals' decisions to avert or mitigate the consequences of environmental deterioration can shed light on how people value environmental quality (averting behavior estimates). In other cases, individuals reveal their preferences for environmental goods in the housing market (hedonic property value methods), or for related health risks in labor markets (hedonic wage methods). In still other cases, individuals reveal their demand for recreational amenities through their decisions to travel to specific locations

[27] For a recent survey of computable general equilibrium models, see Conrad (2002), and for an application of CGE modeling to estimate the costs of the Clean Air and Clean Water Acts, see Hazilla and Kopp (1990).

[28] Three of the *ex ante* cost analysis were underestimates; the other eleven were approximately correct.

[29] On this, also see: Hammitt (2000).

[30] For an intellectual history of developments in this area, see Cropper (2000), and for a survey of theoretical underpinnings and empirical issues associated with alternative benefit estimation methods, see Freeman (2003).

(Hotelling–Clawson–Knetsch and related recreation-demand methods). In addition, empirical evidence of environmental benefits may be obtained when individuals express their willingness to pay for a privately-traded option to use a freely-available public good. This set of revealed preference methods can be used to estimate the trade-offs that are at the heart of environmental valuation, but—as explained below—the scope of potential application of these methods is limited.

The *averting behavior method*, in which values of willingness to pay are inferred from observations of people's behavioral responses to changes in environmental quality, is grounded in the household production function framework.[31] People sometimes take actions to reduce the risk (averting behavior) or lessen the impacts (mitigating behavior) of environmental damages, for example, by purchasing water filters or bottled water. In theory, people's perceptions of the cost of averting behavior and its effectiveness should be measured (Cropper and Freeman, 1991), but in practice actual expenditures on averting and mitigating behaviors are typically employed, with the results sometimes interpreted as constituting a lower bound on willingness to pay. Such an interpretation, however, can be misleading (Shogren and Crocker, 1991, 1999). An additional challenge is posed by the necessity of disentangling attributes of the market good or service. For example, bottled water may be safer, taste better, and be more convenient. In this case, willingness to pay for safer water might be overestimated by an averting behavior approach. On the other hand, since bicycle helmets are uncomfortable, expenditures on such equipment could lead to an underestimate of willingness to pay for risk reduction.

Hedonic pricing methods are founded on the proposition that people value goods in terms of the bundles of attributes that constitute those goods. In theory, the value of the environmental component of a particular good can be extracted by statistically decomposing the value of the total good into willingness to pay for multiple attributes.[32] In the environmental sphere, hedonic methods have been applied to property values and to wages.

Hedonic property value methods employ data on residential property values and home characteristics, including structural, neighborhood, and environmental quality attributes.[33] By regressing the property value on key attributes, the hedonic price function

[31] See Becker (1965) for the early development of the household production method, and Bockstael and McConnell (1983) for the conditions under which the benefits of a public good can be inferred from the demand function for a related private, market good. Mäler's (1985) theoretical explication builds upon a proposal by Ridker (1967). For an early application to human health, see Grossman (1972), and for a complete theoretical exposition, see Courant and Porter (1981).

[32] The hedonic pricing method was originated by Waugh (1928), but it was Court (1939) who developed the method using multiple regression techniques. Hedonic pricing was revived and further developed econometrically by Griliches (1961). Most applications in the environmental realm stem from Rosen (1974). For an examination of the conditions under which the results from the hedonic price function can be used for benefit estimation, see: Bartik (1988a). More recent treatments include those by Ekeland, Heckman, and Nesheim (2002, 2004), which reflect on the identification issues originally addressed by Brown and Rosen (1982). Surveys are provided by Palmquist (1991) and Taylor (2003).

[33] For surveys of methodological developments and applications of hedonic property methods, see: Bartik and Smith (1987); and Palmquist (2005).

is estimated:

$$P = f(\underset{\sim}{x}, \underset{\sim}{z}, e) \tag{3}$$

where $P =$ housing price (includes land);
$\underset{\sim}{x} =$ vector of structural attributes;
$\underset{\sim}{z} =$ vector of neighborhood attributes; and
$e =$ environmental attribute of concern.

From the estimated hedonic price function of equation (3), the marginal implicit price of any attribute, including environmental quality, can be calculated as the partial derivative of the housing price with respect to the given attribute:

$$\frac{\partial P}{\partial e} = \frac{\partial f(\cdot)}{\partial e} = P_e \tag{4}$$

This marginal implicit price, P_e, measures the aggregate marginal willingness to pay for the attribute in question, and it may be interesting in and of itself. For purposes of benefit estimation, however, the demand function for the attribute is required, and so it becomes necessary to examine how the marginal implicit price of the environmental attribute calculated from equation (4) varies with changes in the quantity of the attribute and other relevant variables. If the hedonic price equation (3) is non-linear, then fitted values of P_e can be calculated as e is varied, and a second-stage equation can be estimated:

$$\hat{P}_e = g(e, \underset{\sim}{y}) \tag{5}$$

where $\hat{P}_e =$ the fitted value of the marginal implicit price of e from the first-stage equation; and
$\underset{\sim}{y} =$ a vector of factors that affect marginal willingness to pay for e, including buyer characteristics.

Equation (5), above, has been interpreted as the demand function for the environmental attribute—from which benefits (consumers surplus) can be estimated in the usual way—but there are several important issues and problems.

Most important among the problems confronting the use of the hedonic property method for environmental benefit estimation is the question of whether a demand function has actually been estimated, since environmental quality may affect both the demand for housing and the supply of housing. Thus, the classic identification problem of econometrics arises. In addition, hedonic property value methods build upon a model of the housing market that is in equilibrium for all attributes, with buyers and sellers having full information. Informational asymmetries may distort the analysis, particularly if perceptions about environmental attributes are different from scientific measurements of these values. That is, if individuals' perceptions of environmental attributes do not correspond to actual measurement of attributes, then estimated marginal implicit prices will be biased (possibly upward, possibly downward, depending upon the nature of perceptions).

Because the hedonic property method is based on analysis of marginal changes, it should not be applied to analysis of policies with large anticipated effects, and because the method's data requirements are considerable, omitted variable bias may be a problem. Finally, although the method seems very well suited for some environmental attributes, including noise abatement and proximity to waste sites, many other environmental amenities do not lend themselves to this type of analysis.

A related benefit-estimation method frequently employed in the environmental policy domain is the *hedonic wage method*, which is based on the empirical reality that individuals in well functioning labor markets make trade-offs between wages and risk of on-the-job injuries (or death). In a hedonic context, a job is a bundle of characteristics, including its wage, responsibilities, and risk, among others factors.[34] Two jobs that require the same skill level but have different risks of on-the-job mortality will pay different wages. On the labor supply side, employees tend to require extra compensation to accept jobs with greater risks; and on the labor demand side, employers are willing to offer higher wages to attract workers to riskier jobs. Hence, labor market data on wages and job characteristics can be used to estimate econometrically people's marginal implicit price of risk, that is, their valuation of risk. By regressing the wage on key attributes, the hedonic price function is estimated:

$$W = h(\underset{\sim}{x}, r) \tag{6}$$

where W = wage (in annual terms);
$\underset{\sim}{x}$ = vector of worker and job characteristics; and
r = mortality risk of job.

The marginal implicit price of risk is calculated as the partial derivative of the annual wage with respect to the measured mortality risk:

$$\frac{\partial W}{\partial r} = \frac{\partial h(\cdot)}{\partial r} = W_r \tag{7}$$

Note that the marginal implicit price of risk is the average annual income necessary to compensate a worker for a marginal change in risk throughout the year. This marginal implicit price varies with the level of risk.

Many of the issues that arise with the hedonic property value method have parallels here. First, there is the possibility of simultaneity: causality between risk and wages can run in both directions. For example, higher ambient air pollution might lead to higher compensating wages, but higher wages and incomes in an area may lead to more automobiles and hence more air pollution. Also, if individuals' perceptions of risk do not correspond with actual risks, then the marginal implicit price of risk calculated from a hedonic wage study will be biased, although, as before, the direction of the bias is not obvious. Imperfections in labor markets (less than perfect mobility) can cause problems,

[34] For a detailed treatment of the hedonic wage model, see Viscusi (1992, 1993).

but more important are the significant data requirements that can lead to omitted variable bias.[35]

Direct applications of the method in the environmental realm would appear to be severely limited. Indeed, direct application is limited to occupational, as opposed to environmental exposures and risks. Yet hedonic wage methods are of considerable importance in the environmental policy realm, because the results from hedonic wage studies have frequently been used through "benefit transfer" to infer the value of a statistical life (VSL), as we discuss below in section 2.1.2.5.5.

Recreational activities represent a potentially large class of benefits that are particularly important in assessing policies affecting the use of public lands. The models used to estimate recreation demand fall within the class of household production models, discussed above. First, *travel cost models* (or Hotelling–Clawson–Knetsch models) use information about time and money spent visiting a site to infer the value of that recreational resource. The simplest version of the method involves one site and uses data from surveys of users from various geographic origins, together with estimates of the cost of travel and opportunity cost of time to infer a demand function relating the number of trips to the site as a function of people's willingness to pay for the experience.[36]

The most significant limitation of the simplest travel-cost model is the omission of potential substitute sites. Although one obvious approach is to include the price (travel and opportunity cost) of substitute sites as additional independent variables, better approaches involve multi-site travel cost models or the use of random utility models. *Random utility models* explicitly model the consumer's decision to choose a particular site from alternative recreation locations, assessing the probability of visiting each location. The most important attribute of random utility models is that they can be used to value changes in environmental quality by comparing decisions to visit alternative sites.[37]

All recreation demand models share a set of limitations. First, the valuation of costs depends critically on empirical estimates of the opportunity cost of (leisure) time, which is notoriously difficult to estimate. Also, most trips to a recreation site are part of a multi-purpose experience. If this is ignored, willingness to pay will be over-estimated. In addition, random utility models rely on people's perceptions of environmental quality changes, and so changes that are difficult to observe may be valued "incorrectly."

[35] If individuals change jobs and homes simultaneously—not an unreasonable expectation in some cases—then the observed marginal willingness to pay will reflect both the labor and property markets. On this, see: Rosen (1979); Roback (1982); and Bartik and Smith (1987).

[36] The conceptual approach was proposed by Harold Hotelling in a 1954 letter to the Director of the U.S. National Park Service, and the method was subsequently developed and applied by Davis (1963) and Clawson and Knetsch (1966). For a survey of travel cost models, see Bockstael (1996); and for a recent survey of recreation demand models, see Phaneuf and Smith (2005).

[37] For detailed treatments of random utility/discrete choice models, see: Bockstael, Hanemann, and Strand (1986); Herriges and Kling (1999); Haab and McConnell (2002); and Parsons (2003).

Finally, like all revealed-preference approaches, recreation demand models can be used to estimate use value only; non-use value cannot be examined.[38]

An alternative approach to assessing people's willingness to pay for recreational experiences is to draw on evidence from *private options to use public goods*. This approach also fits within the household production framework, and is based upon the notion of estimating the derived demand for a privately traded option to utilize a freely-available public good. In particular, the demand for state fishing licenses has been used to infer the benefits of recreational fishing (Snyder, Stavins, and Wagner, 2003). Using panel data on state fishing license sales and prices for the continental United States over a fifteen-year period, combined with data on substitute prices and demographic variables, a license demand function was estimated, from which the expected benefits of a recreational fishing day were derived.

In summary, revealed-preference methods of environmental benefit estimation are based upon sound theoretical foundations and can be empirically effective. If well executed, these methods can produce relatively accurate (that is, unbiased) and relatively precise (that is, low variance) estimates. These approaches are therefore strongly favored by economists, both on theoretical and empirical grounds. But revealed preference methods are severely limited in the scope of their direct applicability. In many situations, it is simply not possible to observe behavior that reveals people's valuations of changes in environmental goods and services. This is particularly true with non-use values. With no standard market trade-offs to observe, economists must resort to surveys in which they construct hypothetical markets, employing stated preference, as opposed to revealed preference methods.

2.1.2.5.2. Stated preference methods In the best known stated preference method, *contingent valuation*, survey respondents are presented with scenarios that require them to trade-off, hypothetically, something for a change in the environmental good or service in question. Stated preference methods depend on directly questioning individuals about their valuation of changes in environmental quality. While controversial because of the potential for biased answers, based on intentions rather than actions, stated preference methods are the only way to estimate non-use values for environmental goods.

Contingent valuation (CV) presents survey subjects with a hypothetical increase or decrease in environmental quality and asks how much they would be willing to pay or accept to enact or prevent such changes. The essential steps in carrying out a CV study are: clearly defining the good or service and the policy-induced change in the good or service to be valued; identifying the geographical scope of the "market;" conducting focus groups on components of the survey; pretesting the survey instrument; administering the survey to a random sample of the market; testing the results for reliability (bias) and validity (theoretical correspondence); and possibly using the elicited

[38] On the possibility of using corner-solution models of recreational behavior to estimate non-use values (employing important assumptions along the way), see: Herriges, Kling, and Phaneuf (2004).

willingness-to-pay data for various quantities of the good/service to construct a demand function, and estimate benefits.[39]

The CV survey instrument itself is used to: collect information on the consumer's past, present, and expected future use of the environmental good (or service); collect information on the respondent's socioeconomic characteristics; present a hypothetical scenario describing a change in the good to be valued; present a specific hypothetical payment vehicle, which is both plausible and understandable (examples include taxes, user fees, and product prices); and elicit the respondent's willingness to pay, reminding the respondent of the existence of substitutes.

Elicitation methods have been of four principal types. First, the simplest approach is to ask people for their maximum willingness to pay, but there are few real markets in which individuals are actually asked to generate their reservation prices, and so this method is considered unreliable. Second, in a bidding game, the researcher begins by stating a willingness-to-pay number, asks for a yes-no response, and then increases or decreases the amount until indifference is achieved. The problem with this approach is the inevitable introduction of significant starting-point bias. Third, a related approach is the use of a payment card to be shown to the respondent, but the problem here is that the range of WTP on the card may still introduce bias, and the approach cannot be used with telephone surveys. Fourth and finally, the referendum (discrete choice) approach is favored by researchers. Here, each respondent is offered a different WTP number, to which a simple yes-no response is solicited. This approach minimizes bias, but requires considerably more observations.

The primary advantage of contingent valuation is that it can be applied to a wide range of situations, including use as well as non-use value, but potential problems remain. First, respondents may not understand what they are being asked to value. This may introduce greater variance, if not bias in responses. Likewise, respondents may not take the hypothetical market seriously, because no budget constraint is actually imposed. This can increase variance and bias. On the other hand, if the scenario is "too realistic," strategic bias may be expected to show up in responses. Finally, the "warm glow effect" may plague some stated preference surveys: people may purchase moral satisfaction with large, but unreal statements of their willingness-to-pay (Andreoni, 1995). For example, in one CV study, it was found that 63 percent of respondents indicated they were willing to pay $30 to a specific leading Norwegian environmental organization to protect resources. But when the same organization followed up with mail solicitations to the same sample, fewer than 10 percent of the original respondents contributed anything (Seip and Strand, 1992).

The 1989 Exxon Valdez oil spill in Prince William Sound off the coast of Alaska led to massive litigation, and—as a consequence—resulted in the most prominent use

[39] For a comprehensive treatment of contingent valuation methods, see Mitchell and Carson (1989). For more recent surveys, see: Brown (2003); and Boyle (2003).

ever of the concept of non-use value and the method of contingent valuation for its estimation.[40] The result was a symposium sponsored by Exxon attacking the CV method (Hausman, 1993), and the creation of a government panel—established by the National Oceanic and Atmospheric Administration (NOAA) and chaired by two Nobel laureates in economics—to assess the scientific validity of the CV method (Arrow et al., 1993). The NOAA panel concluded that "CV studies can produce estimates reliable enough to be the starting point of a judicial process of damage assessment, including lost passive (non-use) values" (Arrow et al., 1993, p. 4610). The panel offered its approval of CV methods subject to a set of best-practice guidelines. Since that time, economists have continued to seek ways to improve CV methods and to verify reliability through: replication of CV results; comparison of CV results with other estimates (Hanneman, 1994); and—where possible—comparison of CV results with actual behavior. Nevertheless, some economists remain highly skeptical of this method.[41]

2.1.2.5.3. Fallacious methods of "benefit estimation" It is important to distinguish the averting behavior method, described above, from so-called "avoided-cost measures of benefits" in general, which are attempts to substitute for real measures of benefits the cost of the next most costly means of achieving some goal. Unless individuals have demonstrated their willingness to undertake voluntarily the alternative activities—as in the case of averting behavior methods—using costs as proxies for benefits is illegitimate; it simply converts what would be a benefit–cost comparison into a cost-cost (that is, cost-effectiveness) comparison. By applying "avoided-cost measures of benefits," any proposed project can be made to appear desirable. By taking the next most costly approach of achieving an objective and calling that the project's benefits, one will always find that "benefits"—so measured—exceed costs.

Related to attempts to substitute costs for true measure of benefits is the so-called "societal revealed preference" (SRP) approach, whereby analysts seek to infer the benefits of a proposed policy from the costs of previous regulatory actions. Of course, true revealed preference benefit estimation methods require that individuals or groups voluntarily undertake actions and pay the costs of undertaking those actions. The SRP method fails this test. Only if the previous regulation itself passed a benefit–cost test could the costs of that regulation possibly be assumed to have any particular relation to its benefits. The SRP method is not a revealed-preference method, and indeed is not a benefit-estimation method at all, but—at most—a cost-effectiveness comparison. The purpose of a benefit–cost analysis is to assess policies by contrasting their benefits and their costs; the SRP approach reverses this, taking the fact that a policy exists as evidence that its benefits exceed its costs (and therefore that its benefits can be proxied by

[40] The CV study carried out to estimate non-use value lost as a result of the Exxon Valdez oil spill was eventually published (Carson et al., 2003).

[41] See Portney (1994) for an overview of the debate, Diamond and Hausman (1994) for a skeptical view, and Hanneman (1994) for a defense of CV methods. More recent contributions include: Carson, Flores, and Hanemann (1998); Cameron et al. (2002); and Champ, Boyle, and Brown (2003).

its costs, at a minimum). The use of such approaches would stand the very process of regulatory impact analysis on its head.

Finally, an approach frequently used by government agencies and others in attempts to value changes in morbidity (non-fatal health effects) is the so-called *cost of illness* method (U.S. Environmental Protection Agency, 2002). This approach does not provide a theoretically correct measure of willingness to pay or willingness to accept, but instead measures explicit market costs resulting from changes in the incidence of illness. Direct and indirect medical costs are included, where direct costs refer to diagnosis, treatment, and rehabilitation, and indirect costs refer to the loss of productivity attributable to illness. "Pain and suffering" and averting behavior are not included. Cost-of-illness estimates have therefore been interpreted as providing a lower bound on willingness to pay (Harrington and Portney, 1987), but this may not be the case because the reality of individuals passing costs on to third parties (insurers, hospitals, and employers) means that costs incurred may overstate true individual willingness to pay.

2.1.2.5.4. Benefit transfer Because of the considerable time and cost of both revealed-preference and stated-preference valuation methods, government agencies—including the U.S. Environmental Protection Agency—frequently rely on the transfer of existing estimates from previous research (the "study case") to new contexts (the "policy case"). Such benefit-transfers are very inexpensive compared with original research, but the estimates are far less accurate and reliable, and inevitably introduce arbitrary elements of judgment into the analysis.

Three principal benefit transfer methods have been employed. First, point estimates involve the simple adoption of a benefit number from a previous study. This approach is generally considered unacceptable. Second, a benefit function may be adopted from the study case, plus values of exogenous variables from the policy case; then benefits can be estimated. Third, if such benefit functions are not available from previous research, a meta-analysis may be carried out, combining values from a variety of previous studies, estimating a statistical relationship of the factors affecting benefits, and then employing values of exogenous variable from the policy case in order to estimate (the fitted value of) benefits (U.S. Environmental Protection Agency, 1993, 2000a).

Two major criteria are useful for judging benefit-transfer exercises (Desvousges, Johnson, and Banzhaf, 1998). The first criterion is soundness—was the study-case analysis itself of sufficient quality? The second criterion is similarity. The basic commodities analyzed in the study case and the policy case should be essentially equivalent; the baselines and the degrees of change in the environmental good or service should be similar; and the affected populations should be similar. This is particularly challenging in the natural resources context, because values tend to be highly dependent upon location, suggesting the infeasibility of meeting the similarity condition (Rosenberger and Loomis, 2003).

2.1.2.5.5. Valuing mortality risk reductions How much would individuals sacrifice to achieve a small reduction in the probability of death during a given period of time? How much compensation would individuals require to accept a small increase in that probability? These are reasonable economic questions, given the fact that most environmental regulatory programs result in small changes in individuals' mortality risks. Empirical methods, considered above, including hedonic wages studies, averted behavior, and contingent valuation, can provide estimates of marginal willingness to pay or willingness to accept for small changes in mortality risk. For purposes of benefit transfer, such estimates have been normalized into measures of the "value of a statistical life" (VSL).[42]

The VSL is *not* the value of an individual life—neither in ethical terms, nor in technical, economic terms. Rather it is simply a convention:[43]

$$\mathrm{VSL} = \frac{\text{MWTP or MWTA (from hedonic wage or CV)}}{\text{Small Risk Change}} \tag{8}$$

where MWTP and MWTA, respectively, refer to marginal willingness to pay and marginal willingness to accept. For example, if people are willing, on average, to pay \$12 for a risk reduction from 5 in 500,000 to 4 in 500,000, equation (8) would yield:

$$\mathrm{VSL} = \frac{\$12}{0.000002} = \$6{,}000{,}000 \tag{9}$$

Thus, VSL quantifies the aggregate amount that a group of individuals are willing to pay for small reductions in risk, standardized (extrapolated) for a risk change of 1.0.[44] It is not the economic value of an individual life. The VSL calculation above does not signify that an individual would pay \$6 million to avoid (certain) death this year, or accept (certain) death this year in exchange for \$6 million. It does imply that 100,000 similar people would together pay \$6 million to eliminate the risk that is expected to kill one of them randomly this year.[45]

There has been considerable debate regarding whether and how VSLs should be adjusted for risk characteristics, including the latency periods of pre-mortality illness, the dread associated with some forms of mortality, and the difference between voluntary and involuntary risk. Discounting has been the usual way of handling any latency period prior to mortality, but this may oversimplify how individuals value future impacts

[42] For comprehensive surveys of the VSL literature, see: Fisher, Chestnut, and Violette (1989), Miller (1989), and Viscusi (1993).

[43] The "convention" is to express the marginal willingness to pay for a small reduction in mortality risk or marginal willingness to accept compensation for a small increase in mortality risk, normalized for a risk change of 1.0. It is critical to understand that the convention could just as easily be for a risk change of one in a million. Indeed, if that were the convention, the usefulness of the device for benefit analysis would not be affected in the least, the unfortunate and misleading name of "value of a statistical life" would be avoided, and much of the ensuing controversy might not have arisen. Unfortunately, we are stuck with the normalization and the name, or at least the abbreviation, VSL.

[44] The first formal development of the concept of willingness to pay for mortality risk reductions was by Jones-Lee (1974).

[45] The U.S. Environmental Protection Agency employs a VSL of \$6.2 million in Regulatory Impact Analyses. This is the average of 26 (21 hedonic wage and 5 CV) studies upon which EPA draws for its calculation (U.S. Environmental Protection Agency, 2000a).

(Horowitz and Carson, 1990; Rowlatt et al., 1998).[46] The dread and pain associated with some forms of mortality is clearly relevant, but is properly considered as a morbidity, not a mortality, effect. Since VSLs draw largely upon hedonic wage studies (see above), they reflect valuations of voluntary risk, but their application to environmental policy assessment is related to involuntary risk.

It is also reasonable to ask whether VSLs should be adjusted for population characteristics. Although there is consistent evidence that mortality risk valuation and income (wealth) are highly correlated, evidence on correlation of valuations and health status is mixed.[47] Perhaps most important, it is expected that people's willingness to pay for small changes in risk varies over the course of their lives. But the relationship between age and risk valuation is complicated. Standard economic theory would suggest that younger people would have higher values for risk reduction because they have a longer expected life remaining before them and thus a higher expected lifetime utility (Moore and Viscusi, 1988; Cropper and Sussman, 1990). On the other hand, some models and empirical evidence suggest that older people may in fact have a higher demand for reducing mortality risks than younger people, and that the value of a life may follow an "inverted-U" shape over the life-cycle, with its peak during mid-life (Shepard and Zeckhauser, 1982; Jones-Lee, Hammerton, and Philips, 1985; Ehrlich and Chuma, 1990; Krupnick et al., 2002; Mrozek and Taylor, 2002; Viscusi and Aldy, 2003; Alberini et al., 2004).

Valuations of non-fatal health effects (morbidity) are also required for many benefit–cost analyses in the environmental realm. The theoretically appropriate measure is aggregate willingness-to-pay to reduce the risk of a given health effect, but—as indicated above—cost-of-illness measurements have been used in administrative and judicial contexts when better estimates were not available. Measuring morbidity effects can be more difficult than estimating mortality impacts because of variations in health endpoints.[48]

2.1.2.6. Critiques of benefit–cost analysis In addition to criticism (discussed above in section 2.1.1) of the Kaldor–Hicks criterion as a decision rule, there has been considerable criticism of the use of benefit–cost analysis in the environmental realm, both on conceptual and empirical grounds.[49] The most common conceptual objection to benefit–cost analysis from non-economists is that monetary estimates of environmental quality are impossible and/or unethical. Some have argued that the environment has an intrinsic

[46] Revesz (1999) argues that discounting is ethically unjustified when considering harm to future generations. Cropper and Sussman (1990) identify an alternate method that is more theoretically sound than simple discount rates, but data-intensive. Slovic (1987) provides a review of risk perception issues.

[47] See, for example, Desvousges et al. (1996).

[48] For general references concerned with valuing morbidity, including published estimates of the valuation of many specific effects, see: Tolley, Kenkel, and Fabian (1994), Johansson (1995), Cropper (2000), and U.S. Environmental Protection Agency (2000b, 2002).

[49] See Adler and Posner (2001) for a collection of critiques of benefit–cost analysis and responses to the critiques. Sen (2000) provides a detailed discussion of the full set of conceptual assumptions required for benefit–cost analysis. Also see Kaplow and Shavell (2002a) and the discussion in section 3.1.3.3, below.

value that cannot be quantified numerically, or that attaching a monetary value to environmental quality is ethically wrong because environmental quality should be treated as a basic right that must be protected, regardless of whether the benefits outweigh the costs (Kelman, 1981a). Of course, economic value has a very specific definition: it is a measure of those things that people would be willing to give up to have environmental quality, whether or not it is traded in markets. The implementation of all rights, including those held to be fundamental, requires real resources and imposes real costs. Adding information to the process through benefit–cost analysis can serve to improve decision-making.[50]

More recently, some critics have questioned the empirical methods used for valuing marginal willingness to pay to avoid and willingness to accept compensation to endure incremental changes in risk of mortality (and morbidity). Unfortunately, some of the most prominent critiques have been premised upon fundamental misunderstanding of those same theories and empirical methods, and have been based upon misleading straw-man caricatures of the positions of economists (Heinzerling, 1998; Ackerman and Heinzerling, 2002).

In this context, although formal benefit–cost analysis should not be viewed as either necessary or sufficient for designing sensible public policy, it can provide an exceptionally useful framework for consistently organizing disparate information, and in this way, it can greatly improve the process and hence the outcome of policy analysis (Arrow et al., 1996b). Economists share concerns about the empirical reliability of benefit–cost estimation methods in specific applications, as highlighted throughout our discussion above. More broadly, economists recognize that while benefit–cost analysis can be very helpful to decision makers, it ought to be considered as an aid to decision makers, not a substitute for decision making.

2.1.3. Other approaches to analyzing the goals of environmental policies

Decision-makers and scholars have proposed other evaluation criteria with which environmental policies might be assessed. One approach, reflected in prevailing interpretations of the Clean Air Act and some other environmental laws, has been to claim to rely solely on biophysical (that is, natural science) information to identify policies that eliminate environmental risks altogether or reduce them to levels deemed acceptable. The Clean Air Act, for example, has been construed to adopt this approach, directing EPA to set its ambient air quality standards at levels that will protect public health with an adequate margin of safety (see section 2.2.2, below). Some legal scholars have defended this view (Heinzerling, 2001), but since many environmental pollutants fail to exhibit clear thresholds below which they pose no health effects, such an approach is unworkable as a normative basis for setting environmental standards. More fundamentally, natural science *alone* cannot provide a *normative* basis for setting environmental

[50] A brief and pragmatic defense of the use of benefit–cost analysis is provided by Arrow et al. (1996b). Replies to Kelman's (1981a) critique are provided by DeLong (1981) and Solow (1981).

standards (Coglianese and Marchant, 2004), despite the fact that input from the natural sciences is necessary for implementing economic or most other criteria.

Another problem with a simple risk-elimination approach is that environmental policies can increase certain risks at the same time as they reduce other risks. This motivated the development of *risk-risk analysis* (Lave, 1981), in which health outcomes of alternative policies are calculated and presented directly, without monetary valuation. An important aspect of the analysis is taking into account both the positive, intended effects on health of the policy under consideration and the negative effects that the policy may bring about. Thus, the analysis compares risk reductions caused by a policy with risks created by the policy. For example, a policy that requires power plants to install pollution abatement equipment may reduce the risk of illness due to environmental pollutants, but increase the risk of on the job injury because of construction needed to meet the standards (Lave, 1981).

Clearly risk-risk analysis cannot be used to ascertain whether a policy fulfills the efficiency criteria, because the only costs counted are other health risks; the real resource costs and opportunity costs of implementing the program are ignored. Furthermore, without a common numeraire, policy-makers have no clear standard for comparing different types of health impacts, and so policies cannot be ranked.[51] It has been argued that risk-risk analysis is also flawed because it focuses on negative secondary effects of regulation (ancillary risks), ignoring ancillary benefits (Rascoff and Revesz, 2002). Risk-risk analysis has seen only limited use.

Health-health analysis goes one step further by attempting to quantify resource and opportunity costs, premised on the notion that spending for regulatory programs diverts resources from individuals, causing them to spend less on safety and healthcare, and thereby increasing their morbidity and/or mortality risks.[52] Thus, the public health benefits of a program are contrasted with the negative health effects of the program. A common accounting unit is required, typically the number of fatalities. Health-health analysis provides a measure of "net benefits" (lives saved), but this analytical method suffers from a number of severe limitations (Portney and Stavins, 1994): it does not include other benefits besides saved fatalities; the relatively small cost of environmental regulation as a percentage of individual budgets means that there may be no observable effects on individual health expenditures; and accurate analysis depends on the difficult task of estimating the complex empirical relationship between marginal income changes and health risks.

Distributional analysis provides another approach to analyzing the goals of environmental policies in economic terms. Benefit–cost analysis focuses exclusively on aggregate net benefits, and does not take into account the distributional consequences

[51] See Viscusi, Magat, and Huber (1991), and Graham and Wiener (1995).

[52] Wildavsky (1980) was one of the first to describe the relationship between regulation and increased morbidity or mortality due to loss of disposable income. For empirical analysis, see Keeney (1990, 1997). Lutter and Morrall (1994) provide a theoretical development and a review of the literature. Hahn, Lutter, and Viscusi (2000) provide an empirical evaluation of several regulations using health-health analysis.

of policies. Distributional issues arise, however, on both the benefit and cost sides of the ledger, and appear along a number of dimensions, including: cross-sectional (such as geographic, income, race, sector, and firm characteristics) and intertemporal (such as seasonal, annual, long term, and inter-generational). Distributional equity may be an important societal consideration, particularly in regard to impacts on people of different incomes, and two possible approaches to this issue deserve mention: distributionally weighted benefit–cost analysis and separate distributional analyses.

It is at least theoretically possible to incorporate distributional considerations into benefit–cost analysis by using a system of *distributional weights*,[53] whereby greater attention is given in the analysis to the dollars received or expended by various groups in a benefit–cost analysis. This requires the specification of a set of weights, and there is neither a theoretical nor a political consensus on an appropriate set of weights.[54] Most economists, however, do not advocate attempting to incorporate distributional considerations into benefit–cost analysis (such as via distributional weights), but recommend using separate distributional analysis as a supplement to standard benefit–cost analysis.[55] Such distributional analysis can examine impacts on sub-groups of the population, as well as on the national distribution of income or wealth. Sub-populations that are frequently considered in policy contexts include economic sectors, government, consumers, the elderly, and children. Distributional analysis may also report on potential changes in profitability of firms, changes in employment, plant closures, changes in government revenues, and industry competitiveness.

2.2. Positive issues and analysis

Given the welfare improvements that employment of the efficiency criterion and the related assessment method of benefit–cost analysis could presumably bring to environmental policy, it is reasonable to ask what the reception has been within the three branches of the federal government—executive, legislative, and judicial—to the use of these analytical tools.

2.2.1. Executive orders

At the dawn of the modern environmental movement during the Nixon Administration in the 1970s, the Federal government "placed a high premium on immediate responses to long-neglected problems; emphasized the existence of problems rather than their

[53] For an early treatment of the difficulty of using distributional weights to compare allocations, see Harberger (1978).

[54] Some analyses have used weights based on political behavior such as tax rates. This method was suggested by Eckstein (1961). Applications include Haveman (1965) and Nwaneri (1970).

[55] Examples of applications of distributional analysis to toxic waste contamination include Hird (1993), and Coates, Heid, and Munger (1994), and to air pollution include Gianessi, Peskin, and Wolff (1979) and Bingham, Anderson, and Cooley (1987).

magnitude; and often based its judgments on moral indignation directed at the behavior of those who created pollution and other risks to safety and health" (Sunstein, 2002). But, subsequently Presidents Carter, Reagan, Bush, Clinton, and Bush all introduced formal processes for reviewing economic implications of major environmental, health, and safety regulations, using Regulatory Impact Analysis. Apparently the Executive Branch, charged with designing and implementing regulations, has seen considerable need to develop a yardstick against which the efficiency of regulatory proposals can be assessed, and benefit–cost analysis has been the yardstick of choice.[56]

President Reagan's Executive Order 12,291 directed executive agencies to submit any major proposed rule to the U.S. Office of Management and Budget (OMB) along with a statement assessing its regulatory impact. The order further directed that, "to the extent permitted by law," administrative agencies were not to regulate if the costs of their regulation outweighed the benefits. Supporters of the approach emphasized that it would help achieve least-cost solutions to policy problems by bringing consistency and rationality to the administrative state, while critics contended that OMB review and benefit–cost analysis were intended not to promote efficient regulation, but simply to roll back regulation (Pildes and Sunstein, 1995).

Throughout the Reagan and Bush Administrations, Regulatory Impact Analyses (RIAs) were required under Reagan Executive Orders 12291 and 12498.[57] President George H.W. Bush created a Council on Competitiveness, chaired by Vice President Quayle, which reviewed the impact on industry of selected regulations.

The Clinton Administration, like its two immediate predecessors, issued an Executive Order requiring benefit–cost analysis of all Federal regulations with expected annual costs greater than $100 million.[58] Shortly after taking office in 1993, Clinton abolished the Council on Competitiveness and revoked both of the Reagan orders, replacing them with EO 12866, *Regulatory Planning and Review*.[59] The Clinton EO was substan-

[56] On the other hand, it should be recognized that the Office of Information and Regulatory Affairs (in the U.S. Office of Management and Budget), which reviews draft regulations and manages the process of receiving Regulatory Impact Analyses from the departments and agencies, was itself established by the Congress (through the Paperwork Reduction Act of 1980).

[57] Executive Order (EO) 12291 required agencies to conduct a regulatory impact analysis for all proposed and final rules that were anticipated to have an effect on the national economy in excess of $100 million. Executive Order 12498 required, in addition, a risk assessment for all proposed and final environmental health and safety regulations. EO 12291 has been called the "foremost development in administrative law of the 1980s" (Morgenstern, 1997). The Reagan EOs were not the first presidential effort at regulatory efficiency, however. President Nixon required a "Quality of Life" review of selected regulations in 1971, and President Ford formalized this process in EO 11281 in 1974. President Carter's EO 12044 required analysis of proposed rules and centralized review by the Regulatory Analysis Review Group. The Administration of President George W. Bush has continued to enforce the RIA requirements of Clinton's EO 12866, rather than issuing a new EO (Graham, 2001).

[58] The threshold is not indexed for inflation and has not been modified over time. We refer to year 2000 dollars, unless we indicate otherwise.

[59] In discussing Clinton's EO 12866, many analysts also mention EO 12875, Enhancing the Intergovernmental Partnership, which limited "unfunded mandates". While EO 12875 was part of the Administration's regulatory reform agenda, it did not refer to the efficiency or cost-effectiveness of environmental regulations.

tively and administratively similar to the Reagan orders. It was qualitatively different in tone, however, signaling a less strict efficiency test. While the Reagan orders required that benefits *outweigh* costs, the Clinton order required only that benefits *justify* costs. The Clinton EO allowed that: (1) not all regulatory benefits and costs can be monetized; and (2) non-monetary consequences should be influential in regulatory analysis (Viscusi, 1993). In other ways, the Clinton EO broadened the scope of RIAs to include "distributive impacts" and "equity" in assessing the costs and benefits of particular regulations.[60]

President George W. Bush kept Clinton's executive order in place, further cementing what was already apparent: that the use of benefit–cost analysis in the executive branch has strong bipartisan support. This is not to say, however, that benefit–cost analysis has become a ubiquitous part of all agency decision making. There is evidence that many federal agencies have not complied with the executive orders to engage in meaningful benefit–cost analysis, and the requirements for Regulatory Impact Analysis have not necessarily improved the efficiency of individual Federal environmental rules (Hahn and Dudley, 2004). Further, regulatory impact analysis is required only for major rules, a small fraction of all rules issued by EPA and other agencies. Rules that do not meet this threshold pass under the efficiency radar.

2.2.2. *Legislative enactments*

Over the years, Congress has sent mixed signals regarding the use of benefit–cost analysis in policy evaluation. Some statutes actually require the use of benefit–cost analysis,[61] whereas others have been interpreted to effectively preclude the consideration of benefits and costs in the development of certain regulations.[62] But this has not prevented regulatory agencies from considering the benefits and costs of their regulatory proposals.[63] The problem with such informal, implicit benefit–cost analysis is that it can be

[60] "Costs and benefits shall be understood to include both quantifiable measures (to the fullest extent that these can be usefully estimated) and qualitative measures of costs and benefits that are difficult to quantify, but nevertheless essential to consider. Further, in choosing among alternative regulatory approaches, agencies should select those approaches that maximize net benefits (including potential economic, environmental, public health and safety, and other advantages; distributive impacts; and equity), unless a statute requires another regulatory approach." Executive Order 12866.

[61] Parts of the Clean Water Act, the Consumer Product Safety Act, the Toxic Substances Control Act, the Federal Insecticide, Fungicide, and Rodenticide Act, and the Safe Drinking Water Act explicitly allow or require regulators to consider benefits and costs.

[62] Statutes that have been interpreted (in part, at least) to restrict the ability of regulators to consider benefits and costs include: the Federal Food, Drug, and Cosmetic Act; health standards under the Occupational Safety and Health Act; safety regulations from National Highway and Transportation Safety Agency; the Clean Air Act; the Clean Water Act; the Resource Conservation and Recovery Act; and the Comprehensive Environmental Response, Compensation, and Liability Act.

[63] There is rigorous, empirical evidence that agencies do take into account benefits and costs of regulatory decisions, even when governing statutes do not encourage or allow such analysis to affect decisions. See, for example: Cropper et al. (1992).

unsystematic, not subject to peer review, and carried out behind closed doors, with access limited to the particular friends of the administration. Thus, concerns arise about this approach not only on technical grounds (poor analysis), but on process grounds—it is fundamentally undemocratic.

Despite such arguments, formal benefit–cost analysis has only infrequently been used to help set the stringency of environmental standards. The body politic has favored a very different set of approaches to setting standards, such as that embraced by the Clean Air Act: set the standard to "protect the most sensitive member of the population with an adequate margin of safety." Economists and some legal scholars have spent a great deal of time arguing that such criteria are neither reasonable nor well defined, but little change has occurred.[64]

In the 104th Congress, a major part of the Republicans' "Contract with America" was a regulatory reform bill that would have made meeting a benefit–cost test a *necessary condition* for a broad set of regulatory actions. That bill was narrowly defeated in the Senate, and would have faced a certain Presidential veto, in any case (Sunstein, 1996).[65] Subsequently, Congress considered but did not enact legislation (introduced by former Senator Fred Thompson and Senator Carl Levin) which would have required agencies to conduct (*non-binding*) benefit–cost analyses of new regulations and periodically of existing ones.[66] While this bill never became law, the 106th Congress did pass a major piece of regulatory reform legislation, the Truth in Regulating Act (TIRA), which was signed into law by President Clinton in October 2000. The TIRA established a three-year pilot project beginning in early 2001, which required GAO to review RIAs to evaluate agencies' benefit estimates, cost estimates, and analysis of alternative approaches, upon request by Congress. Because funding was never provided, however, TIRA was not implemented.

In addition to these attempts at cross-cutting regulatory reform, the Congresses of the Clinton years pursued efficiency within specific environmental statutes.[67] In general, Congress was not successful during the 1990s at reforming individual environmen-

[64] The significant and well known heterogeneity of costs per life saved under existing statutes (Table 1) suggests that in the absence of a benefit–cost test aimed at achieving efficiency, much could be accomplished through greater attention to simple cost-effectiveness, that is, achieving given goals or standards at minimum cost. See: Tengs et al. (1995).

[65] But President Clinton did sign the Small Business Regulatory Reform Act of 1996, which provides an opportunity for the Congress to pass legislation that nullifies a regulation that does not pass a benefit–cost test (the nullification itself is then subject to possible Presidential veto, like any act of Congress).

[66] Proposals for the use of a benefit–cost test for setting environmental standards have found a more receptive audience among the states. As of 1996, some 25 of 35 states surveyed reported significant environmental regulatory reform efforts, defined as including the establishment of benefit–cost criteria for promulgation of regulations (Graham and Loevzel, 1997).

[67] During the 1990s, the Congress also pursued reforms of non-environmental statutes that affect environmental regulation. For example, the Accountable Pipeline Safety and Partnership Act of 1996 (104th Congress) requires the Secretary of Transportation to issue pipeline safety regulations only upon justification that benefits exceed costs.

Table 1
Costs of selected environmental, health, and safety regulations that reduce mortality risks

Regulation	Year issued	Agency	Cost per statistical life saved (millions of 2002 dollars)[a]
Logging operations	1994	OSHA	0.1
Unvented space haters	1980	CPSC	0.2
Trihalomethane drinking water standards	1979	EPA	0.3
Food labeling	1993	FDA	0.4
Passive restraints/belts	1984	NHTSA	0.5
Alcohol and drug control	1985	FRA	0.9
Seat cushion flammability	1984	FAA	1.0
Side-impact standards for autos	1990	NHTSA	1.1
Low-altitude windshear equipment and training standards	1988	FAA	1.8
Children’s sleepwear flammability ban	1973	CPSC	2.2
Benzene/fugitive emissions	1984	EPA	3.7
Ethylene dibromide drinking water standard	1991	EPA	6.0
NO_x SIP Call	1998	EPA	6.0
Radionuclides/uranium mines	1984	EPA	6.9
Grain dust	1988	OSHA	11
Methylene chloride	1997	OSHA	13
Arsenic emissions standards for glass plants	1986	EPA	19
Arsenic emissions standards for copper smelters	1986	EPA	27
Hazardous waste listing for petroleum refining sludge	1990	EPA	29
Coke ovens	1976	OSHA	51
Uranium mill tailings (active sites)	1983	EPA	53
Asbestos/construction	1994	OSHA	71
Asbestos ban	1989	EPA	78
Hazardous waste management/wood products	1990	EPA	140
Sewage sludge disposal	1993	EPA	530
Land disposal restrictions/phase II	1994	EPA	2,600
Drinking water/phase II	1992	EPA	19,000
Formaldehyde occupational exposure limit	1987	OSHA	78,000
Solid waste disposal facility criteria	1991	EPA	100,000

[a] Source is Morall (2003). Only final rules are included. Estimates are from respective agencies. Non-mortality and non-health benefits were subtracted from the annual cost (numerator) to generate *net cost*. For each entry, the denominator is the estimated number of statistical lives saved by the regulation annually. Agency abbreviations are as follows. CPSC: Consumer Product Safety Commission; EPA: Environmental Protection Agency; NHTSA: National Highway Traffic Safety Administration; FAA: Federal Aviation Administration; FRA: Federal Railroad Administration; OSHA: Occupational Safety and Health Administration.

tal statutes, although important exceptions were the 1996 Safe Drinking Water Act (SDWA) amendments and the partial reform of pesticide permitting under the Federal Food, Drug and Cosmetic Act (FFDCA).

The 1996 SDWA Amendments include the most far-reaching requirement for benefit–cost analysis in any environmental statute. The Amendments focus EPA regulatory efforts on contaminants that pose the greatest health risks by: (1) requiring benefit–cost analysis of new rules; (2) removing the mandate that EPA regulate 25 new contaminants every three years; (3) allowing EPA to use cost information to adjust its "feasibility standards" for water system reduction of contaminants; and (4) requiring the Administrator to balance risks among contaminants to minimize the overall risk of adverse health effects (Tiemann, 1999). While the Amendments require EPA to determine whether the benefits of each new drinking water maximum contaminant level (MCL) regulation justify the costs, they also allow the Agency to adopt more stringent standards than those that maximize net benefits, explaining the reasons for not selecting the efficient standard.[68]

The Food Quality Protection Act of 1996 amends both the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) and the FFDCA, removing pesticide residues on processed food from the group of Delaney "zero-risk standard" substances. The Delaney standard has long been a target of economic criticism on the grounds that it often leads to associated costs that greatly exceed benefits.[69] While the standard continues to apply to non-pesticide food additives, the Food Quality Protection Act of 1996 eliminated the distinction between pesticide residues on raw foods (which had been regulated under FFDCA section 408) and processed foods (which had been regulated under FFDCA section 409—the Delaney Clause).

It is also important to recognize several failed attempts at changes in individual statutes. Two of the environmental statutes most frequently criticized on efficiency grounds—Superfund (CERCLA) and the Clean Water Act (CWA)—remained relatively untouched by Congress in the 1990s, despite its focus on regulatory reform. Superfund's critics have focused on the low benefits and high costs of achieving the statute's standards (Viscusi, 1992; Breyer, 1993; Hamilton and Viscusi, 1999). Reauthorization and reform were considered during the 105th Congress, but no legislation was passed. Rather than efficiency, distributional aspects of liability issues and questions of how to finance Superfund were the major foci of legislative discussions.[70] The 104th Congress pursued efficiency-oriented reform of the Clean Water Act through the reauthorization process, but the effort failed in the Senate. During the 104th Congress, the House passed a comprehensive Clean Water Act reauthorization (H.R. 961) that would have been more flexible and less prescriptive than the current statute, but the Senate did not take up the bill.

[68] See Safe Drinking Water Act §300g-1 (4)(C). The Amendments do not allow old standards to be subjected to an *ex-post* benefit–cost analysis.

[69] The so-called Delaney clause had the effect of forcing the Food and Drug Administration (FDA) to ban substances from human food supplies that had tested positive as animal carcinogens.

[70] The taxes that support the Superfund trust fund (primarily excise taxes on petroleum and specified chemical feedstocks and a corporate environmental income tax) expired in 1995 and have not been reinstated.

Finally, it is important to note the limited effects of the legislative changes described above. The cross-cutting legislative regulatory reform measures passed in the 1990s and the efficiency-related changes to specific environmental statutes had only limited effects on regulation. This is in part due to differences between the Administration and the Congress in the acceptance of efficiency as an appropriate criterion for managing the environment and natural resources. An additional explanation is the existing statutory bias against benefit–cost analysis in some cases, particularly under the Clean Air Act. In such cases, substantial movement toward efficiency in regulation cannot be expected without substantial changes in the authorizing legislation.

The SDWA Amendments of 1996 incorporated a strong benefit–cost criterion, in comparison with other environmental statutes. However, the decisions made on MCLs since the SDWA Amendments have not placed great weight on the results of required benefit–cost analyses. Two major rules proposed since the 1996 Amendments were those regulating allowable levels of arsenic and radon in drinking water.[71] EPA's benefit–cost analyses for the radon and arsenic MCLs can be interpreted as indicating that monetized costs exceed monetized benefits for both rules, but EPA maintained that the benefits of both rules justify their costs when unquantified benefits are included (U.S. Environmental Protection Agency, 1999).

Likewise, the regulatory reform initiatives passed by Congress in the 1990s apparently did not influence EPA's issuance of National Ambient Air Quality Standards (NAAQS) for ambient ozone and particulate matter in July, 1997. Due to their high potential compliance costs, the revised standards were immediately controversial; both the decision to tighten the standards and the quality of the research used to support the new standards came under fire. EPA's cost estimates for the ozone standard were singled out for criticism (Shogren, 1998; Lutter, 1999). On the other hand, the particulate standard exhibited expected benefits that could well exceed costs by a considerable margin.

The regulated community challenged the new NAAQS in court, and the case reached the U.S. Supreme Court in October, 2000. Under the Clean Air Act, EPA is required to set health-based standards for specified pollutants without consideration of abatement costs. The Supreme Court ruled unanimously in February, 2001, that the Clean Air Act does not allow EPA to consider costs in setting NAAQS for the air pollutants in question (and that the statute's mandate that the NAAQS protect the public health with "an adequate margin of safety" allows an acceptable scope of discretion to EPA).

Overall, the differences in opinion between Congress and the executive branch (especially EPA) on the usefulness of efficiency analysis have resulted in an effective stalemate. Even where statutes have been explicitly altered to require benefit–cost analysis, as was the case for the setting of MCLs under the Safe Drinking Water Act, rules

[71] The arsenic rule was finalized on January 22, 2001, but implementation was delayed while the rule was taken under review by the George W. Bush Administration, citing concerns about the rule's costs and benefits. After an expedited review by the National Academy of Sciences, in October, 2001, EPA Administrator Whitman announced the Agency's intention to enforce the Clinton arsenic standard.

promulgated during the 1990s were not any more or less efficient than rules promulgated during earlier decades. On the other hand, Congressional efforts at generic "regulatory reform" are unlikely to disappear from the policy landscape, and there will continue to be attempts—sometimes successful—to introduce benefit–cost tests into individual environmental statutes.

2.2.3. Judicial recognition

The Federal judiciary also plays a key role in furthering the use of analytical methodologies to assist agency decision making. As Sunstein (2002) notes, over the years courts have implemented a series of benefit–cost "default rules" to deal with Congressional silences and ambiguities. These default rules, while not part of administrative law doctrine, impute to Congressional silence an intent to permit (and perhaps even require) administrative agencies to consider regulatory costs when issuing regulations. The default rules reflect a widespread judicial acceptance of benefit–cost analysis in regulatory rulemaking.[72]

Notably, Justice Stephen Breyer of the U.S. Supreme Court has advocated for more aggressive background rules with respect to risk-risk analysis. In the seminal case of *American Trucking v. Whitman*,[73] referenced above, Justice Breyer argued for a general presumption requiring agencies to engage in benefit–cost analyses when regulating, noting that these analyses would also necessarily involve assessments of risk tradeoffs. In this, he joined ranks with economists and legal scholars who have argued for a judicial presumption in favor of benefit–cost and risk-tradeoff analyses. Breyer's concurrence marks the arrival of risk tradeoff analysis—and health-health tradeoff analysis, in particular—in the Supreme Court and paves the way for future challenges based on such tradeoffs (Rascoff and Revesz, 2002).

2.2.4. A political economy perspective on how standards are set

Granting the merits and relatively widespread acceptance of analytical methods for assessing the tradeoffs inherent in environmental regulation, why has the use of analytical techniques not become more common in environmental policy? Why instead has Congress continued to legislate frequently without regard to benefits and costs? This section reviews positive political economy accounts of how environmental standards are set.

First, some regulations permit established firms to extract rents and establish barriers to entry that convey to them a competitive advantage (Keohane, Revesz, and Stavins, 1998). This is consistent with the empirical reality that the impetus for regulation often comes, either explicitly or implicitly, from regulated firms themselves. For example, a

[72] See: Corrosion Proof Fittings v. EPA, 947 F. 2d. 1201 (1991).

[73] 531 U.S. 457 (2001).

command-and-control standard that limits a firm's aggregate emissions may cause firms to reduce their output to meet the environmental requirement. This output restriction can push the price of a firm's product above its average cost, and as a result, the firm can earn rents. Vigorous competition would dissipate this rent, but environmental regulations often create barriers to entry by imposing stringent pollution standards on new sources, thus giving a significant competitive advantage to established polluters (Maloney and McCormick, 1982).

Second, some industries enjoy strong economies of scale and prefer uniform federal regulation to a patchwork of potentially more efficient state standards. Indeed, having to manufacture different products for sale in different states can destroy the advantages related to economies of scale. National uniformity can come at the expense of regulations more narrowly tailored to achieve optimal outcomes.

Third, even if environmental regulation does not raise the profits of an entire industry, it can benefit certain firms within an industry (Keohane, Revesz, and Stavins, 1998). Firms within an industry likely will incur different costs in the regulatory requirements, because some firms will be able to adjust their production processes more easily than others. These relative beneficiaries of government regulation are thus likely to oppose relaxing regulatory requirements, and may even favor extending them.

Fourth, the impetus for regulation sometimes comes from manufacturers of pollution control equipment, environmentally friendly technologies, or inputs to production processes favored by the regulatory regime. For example, firms specializing in the cleanup of hazardous waste sites emerged in response to the Federal Superfund statute. Similarly, the ethanol industry has strongly supported stricter regulation of gasoline. As a result of its efforts, the Clean Air Act's clean fuels program provides strong incentives for the use of ethanol, and the Federal government has provided large subsidies to ethanol producers.

Fifth, environmental regulation often imposes disproportionate costs on some regions of the country. Regions that incur lower than average costs from regulation become comparatively more attractive to mobile capital, which may bring economic benefits such as jobs and tax revenues. Such regions sometimes push for Federal regulation that will impose disproportionate costs on other regions (Pashigian, 1985).

3. Choosing instruments: the means of environmental policy

Environmental policies typically combine the identification of a goal with some means to achieve that goal. In section 2 of this chapter, we examined the criteria and methods that economics can bring to bear on the choice of targets. In this section, we examine the means—the instruments—that can be employed by governments to achieve given policy objectives. We begin with normative issues and then turn to positive analysis.

3.1. Normative issues and analysis

Even if the goals and targets of environmental policies are taken as given, economic analysis can bring valuable insights to the assessment and design of environmental policies. We begin by considering criteria that can be brought to bear on the search for better policy instruments, and then turn to an enumeration of major categories of environmental policy instruments, including both conventional, command-and-control and the newer breed of market-based instruments. Cross-cutting issues are considered, including uncertainty, technological change, and distributional issues. We examine lessons that emerge from research and experience.

3.1.1. Potential criteria for choosing among policy instruments

A variety of criteria have been posited as relevant for choosing environmental policy instruments, including: (1) will the policy instrument achieve the stated goal or standard; (2) will it do so at the lowest possible cost, including both private-sector compliance and public-sector monitoring and enforcement; (3) will it provide government with the information it needs to implement the policy; (4) will the instrument be flexible in the face of changes in tastes and technology; (5) will the instrument provide dynamic incentives for research, development, and adoption of better pollution-abatement technologies; (6) will the implementation of the policy instrument result in an equitable distribution of the benefits and costs of environmental protection; and (7) will the policy be politically feasible in terms of enactment and implementation? Items (1) through (5) together refer to a comprehensive notion of the criterion of *cost-effectiveness*, while item (6) refers to *distributional equity*, and item (7) refers to *political feasibility*.[74]

First, to be more precise, by cost-effectiveness we mean that allocation of control among sources that results in the aggregate target being achieved at the lowest possible cost, that is, the allocation which satisfies the following cost-minimization problem:

$$\min_{\{r_i\}} C = \sum_{i=1}^{N} c_i(r_i) \tag{10}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} [u_i - r_i] \leq \overline{E} \tag{11}$$

$$\text{and} \quad 0 \leq r_i \leq u_i \tag{12}$$

[74] This list originated with Bohm and Russell (1985). As indicated above, we include the first potential criterion—environmental effectiveness—in a comprehensive definition of cost-effectiveness, but it can also be considered on its own. For example, it has been argued that in some cases the use of market-based instruments has made it politically and/or economically feasible to achieve more stringent goals than otherwise possible (Ellerman et al., 2000; Ellerman, 2003; Harrison, 2003).

where r_i = reductions in emissions (abatement or control) by source i ($i = 1$ to N);
$c_i(r_i)$ = cost function for source i;
C = aggregate cost of control;
u_i = uncontrolled emissions by source i; and
$\overline{E}$ = the aggregate emissions target imposed by the regulatory authority.

If the cost functions are convex, then necessary and sufficient conditions for satisfaction of the constrained optimization problem posed by equations (10) through (12) are the following, among others (Kuhn and Tucker, 1951):

$$\frac{\partial c_i(r_i)}{\partial r_i} - \lambda \geq 0 \tag{13}$$

$$r_i \cdot \left[\frac{\partial c_i(r_i)}{\partial r_i} - \lambda\right] = 0 \tag{14}$$

Equations (13) and (14) together imply the crucial condition for cost-effectiveness that all sources (that exercise some degree of control) experience the same marginal abatement costs (Baumol and Oates, 1988). Thus, when examining alternatives types of environmental policy instruments, a key question is whether particular instruments are likely to result in marginal abatement costs being equated across sources.[75]

3.1.2. Alternative policy instruments

The most frequently employed delineation of environmental policy instruments is that of command-and-control versus market-based approaches. Conventional approaches to regulating the environment—frequently characterized as command-and-control[76]—allow relatively little flexibility in the means of achieving goals. Such policy instruments tend to force firms to take on similar shares of the pollution-control burden, regardless of the cost, sometimes by setting uniform standards for firms, the most prevalent of which are technology- and performance-based standards.[77]

Market-based instruments encourage behavior through market signals, rather than through explicit directives regarding pollution control levels or methods. These policy

[75] For purposes of clarity, the model of cost-effectiveness, above, and subsequent models of specific policy instruments refer to uniformly-mixed flow pollutants. Little additional insight is gained but much is sacrificed in terms of transparency and tractability by modeling more complex non-uniformly mixed stock pollutants. Where the results are not robust to this simplification, we recognize the complexities in the text.

[76] The phrase "command-and-control" is by far the most commonly employed characterization for conventional environmental policy instruments, including uniform performance and technology standards. Admittedly, the phrase has an inescapable negative stigma associated with it, and so a better, more neutral description of this category of policy approaches might be "prescriptive instruments." But because "command-and-control" is the generally accepted name for this category, we employ it in this chapter.

[77] Note that uniform standards can specify the amount of pollution that can be released into the environment (emission standard) or the permissible concentration of pollution in the air, water, or soil (ambient standard). The cost-effective allocation consistent with ambient standards requires equalization of the marginal costs to reduce a unit of ambient concentration, rather than emission.

instruments can reasonably be described as "harnessing market forces,"[78] because if they are well designed and properly implemented, they encourage firms or individuals to undertake pollution control efforts that are in their own interests and that collectively meet policy goals. Market-based instruments fall within four categories: pollution charges, tradeable permits, market-friction reductions, and government subsidy reductions. Liability rules can also be thought of as a market-based instrument, because they provide incentives for firms to take into account the potential environmental damages of their decisions, allowing full flexibility in technology and control practices (Revesz, 1997c).[79]

3.1.2.1. Command-and-control versus market-based instruments Market-based instruments offer the potential for dynamic cost-effectiveness, but problems may arise in translating theory into practice (Hahn and Axtell, 1995), and it has been difficult to measure the magnitude of the gains of moving from command-and-control to incentive-based mechanisms. One frequently-cited survey of eleven empirical studies of air pollution control found that the ratio of actual, aggregate costs of the conventional (command-and-control) approach to the aggregate costs of least-cost benchmarks ranged from 1.07 for sulfate emissions in the Los Angeles area to 22.0 for hydrocarbon emissions at all domestic DuPont plants (Tietenberg, 1985). It is important not to misinterpret these numbers, however, since actual, command-and-control instruments were essentially contrasted with theoretical benchmarks of cost-effectiveness, that is, what a perfectly functioning market-based instrument would achieve in theory.[80] A more useful comparison among policy instruments might involve either idealized versions of both market-based systems and alternatives, or—better yet—realistic versions of both (Hahn and Stavins, 1992).[81]

Where there is significant heterogeneity of costs, command-and-control methods will not be cost-effective. Holding all firms to the same target will be unduly expensive, because it fails to recognize abatement cost heterogeneity. In reality, costs can vary enormously due to production design, physical configuration, age of assets, and other factors. For example, the marginal costs of controlling lead emissions have been estimated to range from $13 to $56,000 per ton (Hartman, Wheeler, and Singh, 1994; Morgenstern, 2000). But where costs are similar among sources, command-and-control

[78] See: Organization for Economic Cooperation and Development (1989, 1991, 1998); and U.S. Environmental Protection Agency (1991, 1992, 2001). Another strain of literature—known as "free market environmentalism"—focuses on the role of private property rights in achieving environmental protection (Anderson and Leal, 1991).

[79] Other taxonomies of regulatory instruments are possible, and some take a more inclusive view, including—for example—contractual approaches. On this, see Menell (2002).

[80] In other cases, researchers have contrasted hypothetical costs of a CAC program with the actual compliance costs associated with the use of a market-based instrument (Keohane, 2003).

[81] Harrington and Morgenstern (2003) attempt to do this by comparing actual experiences in Europe and the United States with market-based and conventional policy instruments.

instruments may perform equivalent to (or better than) market-based instruments, depending on transactions costs, administrative costs, possibilities for strategic behavior, political costs, and the nature of the pollutants (Newell and Stavins, 2003).[82]

In theory, if properly designed and implemented, market-based instruments allow any desired level of pollution cleanup to be realized at the lowest overall cost to society, by providing incentives for the greatest reductions in pollution by those firms that can achieve the reductions most cheaply. Rather than equalizing pollution levels among firms, market-based instruments equalize their marginal abatement costs (Montgomery, 1972; Baumol and Oates, 1988; Tietenberg, 1995). Command-and-control approaches could—in theory—achieve this cost-effective solution, but this would require that different standards be set for each pollution source, and, consequently, that policy makers obtain detailed information about the compliance costs each firm faces. Such information is simply not available to government. By contrast, market-based instruments provide for a cost-effective allocation of the pollution control burden among sources without requiring the government to have this information.

In addition, market-based instruments have the potential to bring down abatement costs over time (that is, to be dynamically cost effective) by providing incentives for companies to adopt cheaper and better pollution-control technologies. This is because with market-based instruments, most clearly with emission taxes, it pays firms to clean up a bit more if a sufficiently low-cost method (technology or process) of doing so can be identified and adopted (Downing and White, 1986; Ellerman, 2003; Maleug, 1989; Milliman and Prince, 1989; Jaffe and Stavins, 1995; Carlson et al., 2000; Popp, 2002; Keohane, 2001; Tietenberg, 2003). However, the ranking among policy instruments, in terms of their respective impacts on technology innovation and diffusion, is not unequivocal (Jaffe, Newell, and Stavins, 2003).

3.1.2.2. Pollution charges

Pollution charge systems assess a fee or tax on the amount of pollution that firms or sources generate (Pigou, 1920). Consequently, it is worthwhile for firms to reduce emissions to the point where their marginal abatement costs are equal to the common tax rate.[83] By definition, actual emissions are equal to unconstrained emissions minus emissions reductions, that is, $e_i = u_i - r_i$. A source's cost minimization problem in the presence of an emissions tax, t, is given by:

$$\min_{\{r_i\}}[c_i(r_i) + t \cdot (u_i - r_i)] \tag{15}$$

$$\text{s.t.} \quad r_i \geq 0 \tag{16}$$

The result for each source is:

$$\frac{\partial c_i(r_i)}{\partial r_i} - t \geq 0 \tag{17}$$

[82] Also see: Atkinson and Lewis (1974); Spofford (1984); and Maloney and Yandle (1984).

[83] For an examination of the robustness of this result in the presence of non-competitive conditions, see Cropper and Oates (1992).

$$r_i \cdot \left[\frac{\partial c_i(r_i)}{\partial r_i} - t \right] = 0 \tag{18}$$

Equations (17) and (18) imply that each source (that exercises a positive level of control) will carry out abatement up to the point where its marginal control costs are equal to the tax rate. Hence, marginal abatement costs will be equated across sources, satisfying the condition for cost-effectiveness specified by equations (13) and (14).

A challenge with charge systems is identifying the appropriate tax rate. For social efficiency, it should be set equal to the marginal benefits of cleanup at the efficient level of cleanup, but policy makers are more likely to think in terms of a desired level of cleanup, and they do not know beforehand how firms will respond to a given level of taxation. An additional problem posed by pollution taxes is associated with their distributional consequences for regulated sources. Despite the fact that such systems minimize aggregate social costs, these systems may be *more* costly than comparable command-and-control instruments *for regulated firms*. This is because with the tax approach, firms pay both their abatement costs *plus* taxes on their residual emissions. For the calculation of aggregate costs in a social benefit–cost or cost-effectiveness analysis, tax payments are simply transfers, and so are excluded from the calculations.

The conventional wisdom is that charge systems have been ignored in the United States, but this is not really correct. If one defines charge systems broadly, a significant number of applications can be identified (Stavins, 2003). The closest that any U.S. charge systems come to operating as true Pigovian taxes may be the increasingly common *unit-charge* systems for financing municipal solid waste collection, where households and businesses are charged the incremental costs of collection and disposal. So-called "pay-as-you-throw" policies, where users pay in proportion to the volume of their waste, are now used in well over one thousand jurisdictions. The collective experience provides evidence that unit charges have been successful in reducing the volume of household waste generated.[84]

Another important set of charge systems implemented in the United States has been *deposit refund systems*, whereby consumers pay a surcharge when purchasing potentially polluting products, and receive a refund when returning the product to an approved center for recycling or proper disposal. A number of states have implemented this approach through "bottle bills" to control litter from beverage containers and to reduce the flow of solid waste to landfills (Bohm, 1981; Menell, 1990), and the concept has also been applied to lead-acid batteries (Table 2).

In addition, there has been considerable use of *environmental user charges* in the United States, through which specific environmentally related services are funded (Table 3). Examples include *insurance premium taxes* (Table 4), such as those formerly used to fund partially the clean-up of hazardous waste sites through the Superfund

[84] See: McFarland (1972); Wertz (1976); Stevens (1978); Efaw and Lanen (1979); Skumatz (1990); Lave and Gruenspecht (1991); Repetto et al. (1992); Miranda et al. (1994); Fullerton and Kinnaman (1996); and Menell (2003).

Table 2
Deposit-refund systems for two regulated products

State	Year of initiation	Amount of deposit ($)
Specified beverage containers		
Oregon	1972	0.05[a]
Vermont	1973	0.05
Maine	1978	0.05
Michigan	1978	0.10
Iowa	1979	0.05
Connecticut	1980	0.05
Delaware	1983	0.05
Massachusetts	1983	0.05
New York	1983	0.05
California	1987	0.025–0.06[b]
Auto batteries		
Minnesota	1988	5.00
Maine	1989	10.00
Rhode Island	1989	5.00
Washington	1989	5.00
Arizona	1990	5.00
Connecticut	1990	5.00
Michigan	1990	6.00
Idaho	1991	5.00
New York	1991	5.00
Wisconsin	1991	5.00
Arkansas	1991	10.00

Source: Stavins (2003).

[a] $0.02 for refillable containers.

[b] Deposits depend upon materials and size of containers.

program (Barthold, 1994).[85] Another set of environmental charges are *sales taxes* on motor fuels, ozone-depleting chemicals, agricultural inputs, and low-mileage motor vehicles (Table 5). Finally, *tax differentiation* has become part of a considerable number of Federal and state attempts to encourage the use of renewable energy sources (Table 6).

3.1.2.3. Tradeable permit systems *Tradeable permits*—in theory—can achieve the same cost-minimizing allocation of the control burden as a charge system,[86] while

[85] The taxes that previously supported the Superfund trust fund—primarily excise taxes on petroleum and specified chemical feedstocks and a corporate environmental income tax—expired in 1995, and have not been reinstated.

[86] Thirty years ago, Crocker (1966) and Dales (1968) independently developed the idea of using transferable discharge permits to allocate the pollution-control burden among sources. Montgomery (1972) provided the

Table 3
Federal user charges

Item taxed	First enacted/ modified	Rate	Use of revenues
Trucks and trailers (excise tax)	1917/1984	12%	Highway Trust Fund/Mass Transit Account
Sport fishing equipment	1917/1984	10% (except 3% for out board motors)	Sport Fishing Restoration Account of Aquatic Resources Trust Fund
Firearms and ammunition	1918/1969	10%	Federal Aid to Wildlife Program
Noncommercial motorboat fuels	1932–1992	$.183/gal	Aquatic Resource Trust Fund
Motor fuels	1932/1993	$.183/gal	Highway Trust Fund/Mass Transit Account
Non-highway recreational fuels & small-engine motor fuels	1932/1993	$.183/gal gasoline $.243/gal diesel	National Recreational Trails Trust Fund and Wetlands Account of Aquatic Resources Trust Fund
Annual use of heavy vehicles	1951/1993	$100–$500/vehicle	Highway Trust Fund/Mass Transit Account
Bows and arrows	1972/1984	11%	Federal Aid to Wildlife Program
Inland waterways fuels	1978/1993	$.233/gal	Inland Waterways Trust Fund

Source: Stavins (2003).

Table 4
Federal insurance premium taxes

Item or action taxed	First enacted/ modified	Rate	Use of revenues
Coal production	1977/1987	$1.10/ton underground; $0.55/ton surface	Black Lung Disability Trust Fund
Chemical production	1980/1986	$0.22 to $4.88/ton	Superfund (CERCLA)
Petroleum production	1980/1986	$0.097/barrel crude	
Corporate income	1986	0.12% of alternative minimum taxable income over $2 million	
Petroleum-based fuels, except propane	1986/1990 (expired 1995)	$.001/gal	Leaking Underground Storage Trust Fund
Petroleum and petroleum products	1989/1990	$.05/barrel	Oil Spill Liability Trust Fund

Source: Stavins (2003).

Table 5
Federal sales taxes

Item or action taxed	First enacted/ modified	Rate	Use of revenues
New tires	1918/1984	$0.15–$0.50/pound	U.S. Treasury
New automobiles exceeding fuel efficiency standards	1978/1990	$1,000–$7,700 per auto	U.S. Treasury
Ozone-depleting substances	1989/1992	$4.35/pound	U.S. Treasury

Source: Stavins (2003).

Table 6
Federal tax differentiation

Item or action taxed	Provision	First enacted/ modified	Rate
Motor fuels excise tax exemptions[a]	Natural gas	1978/1990	$.07/gal
	Methanol	1978/1990	$.06/gal
	Ethanol	1978/1990	$.054/gal
Income tax credits	Alcohol fuels	1980/1990	$.60/gal methanol; $ 0.54/gal ethanol
	Business energy	1980/1990	10% solar; 10% geothermal
	Non-conventional fuels	1980/1990	$3.00/Btu-barrel equivalent of oil
	Wind production	1992	1.5¢/kWh
	Biomass production	1992	1.5¢/kWh
	Electric automobiles	1992	10% credit
Other income tax provisions	Van pools	1978	Tax-free employer provided benefits
	Mass transit passes	1984/1992	Tax-free employer provided benefits
	Utility rebates	1992	Exclusion of subsidies from utilities for energy conservation measures
Tax exempt private activity bonds	Mass transit	1968/1986	Interest exempt from Federal taxation
	Sewage treatment	1968/1986	Interest exempt from Federal taxation
	Solid waste disposal	1968/1986	Interest exempt from Federal taxation
	Waster treatment	1968/1986	Interest exempt from Federal taxation
	High speed rail	1988/1993	Interest exempt from Federal taxation

Source: Stavins (2003).
[a]Exemptions from the motor fuels excise tax of $0.183/gallon (see Table 3).

first rigorous proof that such a system could provide a cost-effective policy instrument. A sizeable literature has followed, much of it stemming from Hahn and Noll (1982). Early surveys were provided by Tietenberg avoiding the problems of uncertain responses by firms and the distributional consequences of taxes.[87] Under a tradable permit system, an allowable overall level of pollution, E, is established, and allocated among firms in the form of permits. Firms that keep their emission levels below their allotted level may sell their surplus permits to other firms or use them to offset excess emissions in other parts of their operations.[88] Let q_{0i} be the initial allocation of emission permits to source i, such that:

$$\sum_{i=1}^{N} q_{0i} = \overline{E} \tag{19}$$

Then, if p is the market-determined price of tradeable permits, a single firm's cost minimization problem is given by:

$$\min_{\{r_i\}}[c_i(r_i) + p \cdot (u_i - r_i - q_{0i})] \tag{20}$$

$$\text{s.t.} \quad r_i \geq 0 \tag{21}$$

The result for each source is:

$$\frac{\partial c_i(r_i)}{\partial r_i} - p \geq 0 \tag{22}$$

$$r_i \cdot \left[\frac{\partial c_i(r_i)}{\partial r_i} - p\right] = 0 \tag{23}$$

Equations (22) and (23) together imply that each source (that exercises a positive level of control) will carry out abatement up to the point where its marginal control costs are equal to the market-determined permit price. Hence, the environmental constraint, E, is satisfied, and marginal abatement costs are equated across sources, satisfying the condition for cost-effectiveness. Note that the unique cost-effective equilibrium is achieved independent of the initial allocation of permits (Montgomery, 1972).[89] This is of great importance politically, as we discuss below in section 3.2.

(1980, 1985). Much of the literature may be traced to Coase's (1960) treatment of negotiated solutions to externality problems. As indicated previously, the simple model posited above, as well as the prior model of emission taxes, assumes the existence of a uniformly-mixed pollutant, in which case the focus of regulation can be exclusively on emissions, as opposed to ambient concentrations. There is a sizable literature that explores tradeable permit and other policy instruments in the context of non-uniformly-mixed pollution problems. See, for example: Montgomery (1972); and Nash and Revesz (2001).

[87] This assumes that the allocation is made without charge, but it could also be through sale or auction, in which case the distributional implications of a comparable tradeable permit program are similar to the emission tax previously described. Likewise, a revenue-neutral emissions tax, in which revenues are refunded to regulated firms (but not in proportion to their emissions levels), can resemble—in distributional terms—a comparable tradeable permit program in which the permits are allocated without charge.

[88] The simple program described above is a "cap-and-trade" system, but some systems operate as "credit programs," where permits or credits are assigned only when a source reduces emissions below what is required by source-specific limits.

[89] This is true unless particularly perverse types of transactions costs are present (Stavins, 1995).

Table 7
Major U.S. tradeable permit systems

Program	Traded commodity	Period of operation	Environmental and economic effects
Emissions Trading Program	Criteria air pollutants under the Clean Air Act	1974–present	Environmental performance unaffected; total savings of \$5–12 billion
Leaded Gasoline Phasedown	Rights for lead in gasoline among refineries	1982–1987	More rapid phaseout of leaded gasoline; \$250 million annual savings
Water Quality Trading	Point-nonpoint sources of nitrogen & phosphorous	1984–1986	No trading occurred, because ambient standards not binding
CFC Trading for Ozone Protection	Production rights for some CFCs, based on depletion potential	1987–present	Environmental targets achieved ahead of schedule; effect of TP system unclear
Heavy Duty Engine Trading	Averaging, banking, and trading of credits for NO_x and particulate emissions	1992–present	Standards achieved; cost savings unknown
Acid Rain Reduction	SO_2 emission allowances; mainly among electric utilities	1995–present	SO_2 reductions achieved ahead of schedule; annual savings of \$1 billion per year
RECLAIM Program	SO_2 and NO_x emissions by large stationary sources	1994–present	Unknown
Northeast Ozone Transport	Primarily NO_x emissions by large stationary sources	1999–present	Unknown

Source: Stavins (2003).

In theory, a number of factors can adversely affect the performance of a tradeable permit system, including: concentration in the permit market (Hahn, 1984; Misolek and Elder, 1989); concentration in the product market (Maleug, 1990); transaction costs (Stavins, 1995); non-profit maximizing behavior, such as sales or staff maximization (Tschirhart, 1984); the preexisting regulatory environment (Bohi and Burtraw, 1992); and the degree of monitoring and enforcement (Keeler, 1991; and Montero, 2003).

Tradeable permits have been the most frequently used market-based system in the United States (U.S. Environmental Protection Agency, 2000a; Tietenberg, 1997a). A selection of programs is summarized in Table 7. The U.S. EPA first experimented with emissions trading in 1974, as part of the Clean Air Act's program for improving local air quality, and later codified these initiatives in its Emissions Trading Program in 1986 (Tietenberg, 1985; Hahn, 1989; Foster and Hahn, 1995). Significant applications include: EPA's emissions trading program (Tietenberg, 1985; Hahn, 1989); the leaded gasoline phasedown; water quality permit trading (Hahn, 1989; Stephenson, Norris, and Shabman, 1998); CFC trading (Hahn and McGartland, 1989); the sulfur dioxide (SO_2) allowance trading system for acid rain control; the RECLAIM program in the Los Angeles metropolitan region (Harrison, 1999); and tradeable development rights for land use.[90] At least two of these programs—lead trading and the SO_2 allowance system—merit further comment.

The purpose of the lead trading program, developed in the 1980s, was to allow gasoline refiners greater flexibility in meeting emission standards at a time when the lead-content of gasoline was reduced to 10 percent of its previous level. In 1982, EPA authorized inter-refinery trading of lead credits, a major purpose of which was to lessen the financial burden on smaller refineries, which were believed to have significantly higher compliance costs. If refiners produced gasoline with a lower lead content than was required, they earned lead credits. In 1985, EPA initiated a program allowing refineries to bank lead credits, and subsequently firms made extensive use of this option. In each year of the program, more than 60 percent of the lead added to gasoline was associated with traded lead credits (Hahn and Hester, 1989), until the program was terminated at the end of 1987, when the lead phasedown was completed.

The lead program was successful in meeting its environmental targets, although it may have produced some temporary geographic shifts in use patterns (Anderson, Hofmann, and Rusin, 1990). Although the benefits of the trading scheme are more difficult to assess, the level of trading activity and the rate at which refiners reduced their production of leaded gasoline suggest that the program was relatively cost-effective (Kerr and Maré, 1997; Nichols, 1997). The high level of trading among firms far surpassed levels observed in earlier environmental markets. EPA estimated savings from the lead trading program of approximately 20 percent below alternative programs that did not provide for lead banking, a cost savings of about $250 million per year (U.S. Environmental Protection Agency, Office of Policy Analysis, 1985). Furthermore, the program appears to have provided measurable incentives for cost-saving technology diffusion (Kerr and Newell, 2003).

The most important application made to date of a market-based instrument for environmental protection has been the SO_2 allowance trading program for acid rain control, established under the Clean Air Act Amendments of 1990, and intended to reduce SO_2 emissions by 10 million tons below 1980 levels (Ferrall, 1991). A robust market of bilateral SO_2 permit trading gradually emerged, resulting in cost savings on the order of $1 billion annually, compared with the costs under some command-and-control regulatory alternatives (Carlson et al., 2000).[91] Although the program had

[90] In addition, the Energy Policy and Conservation Act of 1975 established Corporate Average Fuel Economy (CAFE) standards for automobiles and light trucks, requiring manufacturers to meet minimum sales-weighted average fuel efficiency for their fleets sold in the United States. A penalty is charged per car sold per unit of average fuel efficiency below the standard. The program operates like an intra-firm tradeable permit system, since manufacturers can undertake efficiency improvements wherever they are cheapest within their fleets. For reviews of the program's costs relative to "equivalent" gasoline taxes, see: Crandall et al. (1986); Goldberg (1998); and National Research Council (2002). Light trucks, which are defined by the Federal government to include "sport utility vehicles," face weaker CAFE standards.

[91] The choice of counterfactual for purposes of comparison in such estimates of cost savings is important. The estimate above represents a cost savings of about 30 percent. Employing a different counterfactual for comparison, Keohane (2003) estimates cost savings between 15 and 25 percent.

low levels of trading in its early years (Burtraw, 1996), trading increased significantly over time (Schmalensee et al., 1998; Stavins, 1998; Burtraw and Mansur, 1999; Ellerman et al., 2000).

Concerns were expressed early on that state regulatory authorities would hamper trading in order to protect their domestic coal industries, and some research indicates that state public utility commission cost-recovery rules provided poor guidance for compliance activities (Rose, 1997; Bohi, 1994). Other analysis suggests that this was not a major problem (Bailey, 1996). Similarly, in contrast to early assertions that the structure of EPA's small permit auction market would cause problems (Cason, 1995), the evidence now indicates that this had little or no effect on the vastly more important bilateral trading market (Joskow, Schmalensee, and Bailey, 1998).

The reduction of emissions through the allowance trading program apparently has had exceptionally positive welfare effects, with benefits being as much as six times greater than costs (Burtraw et al., 1998). The large benefits of the program are due mainly to the positive human health impacts of decreased local SO_2 and particulate concentrations, not the ecological impacts of reduced long-distance transport of acid deposition. This contrasts with what was understood and assumed at the time of the program's enactment in 1990.

3.1.2.4. Market friction reduction *Market friction reduction* can also serve as a policy instrument for environmental protection. *Market creation* establishes markets for inputs or outputs associated with environmental quality. Finally, since well-functioning markets depend on the existence of well-informed producers and consumers, *information programs* can help foster market-oriented solutions to environmental problems. *Product labeling requirements* can improve the information set available to consumers, as can various types of *reporting requirements*.

One prominent example of *market creation* is provided by measures that facilitate the voluntary exchange of water rights and thus promote more efficient allocation and use of scarce water supplies (Stavins, 1983; Howe, 1997), and policies that facilitate the restructuring of electricity generation and transmission. The western United States has long been plagued by inefficient use and allocation of its scarce water supplies, largely because users do not have incentives to take actions consistent with economic and environmental values. Economists have noted that federal and state water policies aggravate rather than improve these problems (Anderson, 1983; Frederick, 1986; El-Ashry and Gibbons, 1986; Wahl, 1989). The disparity in water prices over short geographic distances indicates that markets could play a role in solving increasing urban demands for water without the need for new, environmentally-disruptive dams and reservoirs. Reforms have allowed markets in water rights to develop and voluntary exchanges have developed in several states. For example, an agreement was reached to transfer 100,000 acre-feet of water per year from the farmers of the Imperial Irrigation District in southern California to the Metropolitan Water District in the Los Angeles area. Transactions have emerged elsewhere in California, and in Colorado, New Mexico, Arizona, Nevada, and Utah (MacDonnell, 1990).

Table 8
Federal and selected state information programs

Information program	Year of implementation	Enabling legislation
Energy Efficiency Product Labeling	1975	Energy Policy and Conservation Act, Title V
NJ Hazardous Chemical Emissions	1984	New Jersey Community Right-to-Know Act
Toxic Release Inventory	1986	Emergency Planning and Community Right-to-Know Act
CA Hazardous Chemical Emissions	1987	California Air Toxics Hot Spots and Information Assessment Act
CA Proposition 65	1988	California Safe Drinking Water Act and Toxic Enforcement Act
Energy Star	1993	Joint program of the U.S. Environmental Protection Agency and the U.S. Department of Energy

Source: Stavins (2003).

Since well-functioning markets depend, in part, on the existence of well-informed producers and consumers, *information programs* can help foster market-oriented solutions to environmental problems (Table 8).[92] These programs have been of two types. Product labeling requirements have been implemented to improve information sets available to consumers. For example, the U.S. Energy Policy and Conservation Act of 1975 specifies that certain appliances and equipment carry labels with information on products' energy efficiency and estimated energy costs (U.S. Congress, Office of Technology Assessment, 1992). More recently, EPA and the U.S. Department of Energy developed the Energy Star program, in which energy efficient products can display an *EnergyStar* label. And since 1976, the Department of Energy has provided no-cost energy assessments to small and medium-sized manufacturers through its university-based Industrial Assessment Centers (IAC) program. There has been relatively little analysis of the efficacy of such programs, but limited empirical (econometric) evidence suggests that energy-efficiency product labeling has had significant impacts on efficiency improvements, essentially by making consumers and therefore producers more sensitive to energy price changes (Newell, Jaffe, and Stavins, 1999). Also, about half of the projects recommended by assessment teams in the IAC program were subsequently adopted, with firms applying a one to two-year payback period (or about a 50 to 100 percent hurdle rate) to the decisions (Anderson and Newell, 2004).

Another set of information programs has involved reporting requirements. A prominent example is the U.S. Toxics Release Inventory (TRI), established in 1986, which requires firms to make available to the public information on use, storage, and release of

[92] For a comprehensive review of information programs and their apparent efficacy, see Tietenberg (1997b), and for an overview of international experience with "eco-labels," see Morris and Scarlett (1996). Also see Menell (2002).

specific hazardous chemicals. Such information reporting may increase public awareness of firms' actions, and consequent public scrutiny may encourage firms to alter their behavior, although the evidence on outcomes is mixed (U.S. General Accounting Office, 1992; Hamilton, 1995; Konar and Cohen, 1997; Ananathanarayanan, 1998; Hamilton and Viscusi, 1999).

3.1.2.5. Government subsidy reduction *Government subsidy reduction* constitutes another category of market-based instruments. Subsidies are the mirror image of taxes and, in theory, can provide incentives to address environmental problems. Although subsidies can advance environmental quality (see, for example, Jaffe and Stavins, 1995), it is also true that subsidies, in general, have important disadvantages relative to taxes (Dewees and Sims, 1976; Baumol and Oates, 1988). Because subsidies increase profits in an industry, they encourage entry, and can thereby increase industry size and pollution output (Mestelman, 1982; Kohn, 1985).

In practice, rather than internalizing externalities, many subsidies promote economically inefficient and environmentally unsound practices. In such cases, reducing subsidies can increase efficiency and improve environmental quality. For example, because of concerns about global climate change, increased attention has been given to federal subsidies and other programs that promote the use of fossil fuels. An EPA study indicates that eliminating these subsidies would have a significant effect on reducing carbon dioxide (CO_2) emissions (Shelby et al., 1997). The Federal government is involved in the energy sector through the tax system and through a range of individual agency programs. One study indicates that these activities together cost the government $17 billion annually (Koplow, 1993). A substantial share of these U.S. subsidies and programs were enacted during the "oil crises" to encourage the development of domestic energy sources and reduce reliance on imported petroleum. They favor energy supply over energy efficiency.[93] Although there is an economic argument for government policies that encourage new technologies that have particularly high risk or long term payoffs, mature and conventional technologies currently receive nearly 90 percent of the subsidies.[94]

3.1.2.6. Liability rules *Liability rules* have been most frequently employed for acute hazards, particularly for toxic waste sites and for the spill of hazardous materials (Menell, 1991). One important example is the Comprehensive Environmental Response, Compensation, and Liability Act of 1980, which established retroactive liability for companies that are found responsible for the existence of a site requiring clean up.

[93] The Koplow (1993) study claims that end-use efficiency receives $1 from a wide variety of implicit and explicit federal subsidies for every $35 received by energy supply.

[94] On the other hand, federal user charges and insurance premium taxes include significant levies on fossil fuels, and federal tax differentiation has tended to favor renewable energy sources and non-conventional fossil fuels.

Governments can collect cleanup costs and damages from waste producers, waste transporters, handlers, and current and past owners and operators of a site.[95] Similarly, the Oil Pollution Act makes firms liable for cleanup costs, natural resource damages, and third party damages caused by oil spills onto surface waters; and the Clean Water Act makes responsible parties liable for cleanup costs for spills of hazardous substances.

In an *ex post* regulatory scheme,[96] private polluters can be held liable for damages to remedy the harms they cause to an affected individual or group. In theory, the full costs of their polluting activities will thus be internalized, and polluters will reduce the expected harm of their activity up to the point at which further reductions become more costly than the expected liability they face.[97]

The effectiveness of liability rules depends in part on the ability of victims of pollution to bring actions to recover damages. There are five potential problems. First, environmental harms may be widely dispersed, and so the expected payoff may not justify the cost to an individual victim of bringing a lawsuit (Cropper and Oates, 1992). This collective-action problem can partially be addressed by permitting individuals to bring class actions on behalf of all those harmed by polluters. Second, frequently there are many sources of a given pollutant, and hence the aggrieved party (or parties) may not be able to identify the actual source of the damages. Third, many pollution harms have long latency periods, meaning that by the time the harm has manifested itself, actions are barred because of statutes of limitations. In some jurisdictions, however, such statutes begin to run only with the discovery of the harm, not the imposition of the risk. Fourth, many environmental impacts, such as induced disease, are *stochastic* by nature, that is, environmental exposure increases the *probability* of morbidity or mortality. In such cases, it is difficult or impossible to determine with certainty the source of environmental harm. Evidentiary rules that require "a preponderance of the evidence" showing that the plaintiff caused the defendant's harm would not allow recovery under these circumstances. Fifth, a polluter may not have sufficient solvency to pay a large damage award, and the difference between the polluter's total solvency and the full damages will be externalized onto the public.[98]

Nevertheless, liability rules have a central role to play in environmental regulation, because other regulatory tools give rise to their own sets of problems. There are important choices that need to be made in designing liability rules, however. Should polluters

[95] For economic analyses of the Superfund program, see, for example: Hamilton (1993); Gupta, Van Houtven, and Cropper (1996); and Hamilton and Viscusi (1999).

[96] *Ex post* and *ex ante* legal regimes transmit different incentives to private actors. Shavell (1984) argued that the choice between the two regimes should be considered in light of four factors. First, a liability regime might be preferable if private parties have better information than a regulating authority regarding the risks of productive activities. Second, the greater the likelihood that a private party will not be able to pay fully for a harm, the more attractive is a regulatory regime. Third, the greater the chance that private parties will not face the threat of a lawsuit, the more should regulation be favored. Fourth, the administrative costs associated with the two regimes generally weigh in favor of a liability scheme. Also see: Kolstad, Ulen, and Johnson (1990).

[97] This section draws upon Kornhauser and Revesz (2000).

[98] Further, a liability scheme may give private actors an incentive to shed their solvency (through dividends to their shareholders, for example) in order to avoid paying large awards (Kornhauser and Revesz, 1990).

be held jointly and severally liable for the harm they cause? Non-jointly liable? Is a negligence rule preferable to a strict liability rule, or vice-versa? Moreover, which parties should be held liable? Polluters? Site owners? Alternative liability regimes transmit different sets of incentives to private actors and can have dramatically different effects.

3.1.2.6.1. Joint and several versus non-joint liability When a plaintiff's injury results from the actions of multiple parties, the choice between joint and several and non-joint liability arises. Under joint and several liability, the plaintiff can recover the full amount of damages from any one of the defendants who share responsibility for the damage. Under a system of non-joint liability, a plaintiff can only recover from a defendant the share of damages attributable to that defendant.

Several choices must be made with respect to any joint and several liability regime. First, joint and several liability regimes may or may not allow for contribution, whereby a defendant that has paid a disproportionately large share of a particular damage award will be compensated by parties that have paid disproportionately small shares of that award. Second, contribution shares can be determined either by reference to comparative fault or on a *pro rata* basis. Third, in the event that a plaintiff settles with one defendant, the regime must specify by how much the total damage award against the remaining defendants ought to be reduced ("set-off"). Under a "pro tanto set-off rule," the plaintiff's claim against the non-settling defendant is reduced by the amount of the settlement. In contrast, under an apportioned or proportional share set-off rule, the plaintiff's claim against the non-settling defendant is reduced by the share of the liability attributable to the settling defendant. Fourth, when one defendant settles and another defendant litigates and loses, the regime must specify whether, under the pro tanto set-off rule, the settling defendant is protected against contribution actions. Fifth, the legal regime must also indicate whether settling defendants can bring actions for contribution against defendants who settle for less than their share of liability.

Sixth, joint-and-several regimes sometimes protect non-settling defendants from a plaintiff's inadequately low settlements with other defendants through a "good faith" hearing on the settlement's adequacy. And seventh, the regime must specify whether a sued defendant can join a third-party defendant that the plaintiff has declined to name. These choices among rules can have significant impacts on deterrence (Kornhauser and Revesz, 1989, 1990), as well as on the likelihood of inducing settlements.[99]

3.1.2.6.2. Liability extension On whom is liability imposed? Assume that there are two groups of actors: waste generators and disposal site owners. One or both could potentially be held liable for problems associated with the disposal of waste. What liability scheme would be preferable on the grounds of efficiency and deterrence?

A legal regime might impose liability solely on the owner of a hazardous waste site and refuse to extend liability to the generators of that waste. Site owners will, under this

[99] The impact of the possibility of settlement on the choice-of-regime analysis is analyzed in Kornhauser and Revesz (1994a, 1994b).

regime, bear the full costs of the hazardous waste they receive. In a competitive market, disposal site owners will tend to charge generators the marginal cost of disposal. Site owners will have incentives to accept only waste that can properly be disposed of, based on geologic conditions of the site, possible interactions between different types of waste, and other relevant factors. To reduce their disposal costs, generators may seek to reduce the quantity of hazardous waste they produce, and will send their hazardous waste to sites that can process the wastes most effectively—and hence to the sites where wastes will cause the least harm.

While theoretically sound, such full internalization of the site owner's costs is unlikely in practice. Cleanup costs for hazardous waste sites can be extraordinarily large,[100] and site owners will likely not be sufficiently solvent to pay total cleanup costs in the event of a high-cost problem. If the probability of such an event occurring is not zero, if the site owner's solvency is less than the full costs associated with that event, and if the site owner does not fully insure against the risk, then the site owner will not bear the full cost, and will charge a price that will not reflect the full cost of remediation (Shavell, 1987; Pitchford, 1995).

There are two possible solutions to the problem of insolvency (Shavell, 2005). First, polluters could be asked to post a bond equal to possible remediation costs. Given the large costs of environmental clean-up, however, such bonds may drive potential polluters out of the market. Second, polluters can be required to carry insurance for potential liabilities.[101] Because insurance companies' monitoring costs are likely to be high, however, they will only be able to issue insurance based on easily observable factors unrelated to whether the polluter is taking due care to reduce its pollution. Hence, the polluter's premiums will not be reduced if it takes due care, and a significant moral hazard problem arises. Insurance may therefore be unavailable in the environmental context (Abraham, 1988). Moreover, minimum asset requirements could have socially undesirable effects by banning from the activity actors that derive benefits that are higher than the harms they imposed, even in light of their reduced incentives to take care caused by their limited solvency (Shavell, 2005).

As an alternative, liability could be extended only to the generators of hazardous waste, so that the generators bear the full cleanup costs associated with their waste production. In this case, generators could achieve efficient disposal costs in one of two ways. First, generators could shift liability onto site owners by offering them a payment in exchange for an indemnification agreement. This solution is, as discussed above, hampered by the problem of insolvency. Alternatively, to coordinate efficiently and keep their liability to a minimum, generators could contract among themselves to dispose of specified waste at specified locations. There are two difficulties with this approach. First, transaction costs are likely to be prohibitively large, and generators are therefore

[100] The average cleanup cost for a site on the Superfund National Priorities List is $30 million, with many exceeding $100 million.

[101] A voluntary insurance program will prove inadequate because low-solvency polluters will have no incentive to purchase insurance for a cost they will never bear.

likely to act in a non-cooperative manner (Kornhauser and Revesz, 1994a). Moreover, whether generators are subject to either joint and several or non-joint liability, a strict liability regime applied to a group of generators produces under-deterrence (Kornhauser and Revesz, 1989, 1995).

The second problem relates, again, to solvency. Low-solvency generators have less incentive than high-solvency generators to produce an optimal amount of waste. One would expect that high-solvency generators will thus refuse to contract with their low-solvency counterparts, particularly under joint and several liability regimes, where high-solvency generators may be held liable for the full amount of damages caused by low-solvency generators. This distortion of the contracting patterns can reduce welfare (Boyd and Ingberman, 1997). A non-joint liability regime produces the same result if the damages at the site are allocated among generators proportionally to the amount of waste dumped, and the damage function is convex. In that scenario, one generator's decision to dump more than the optimal amount results in higher liability for the other generators (Kornhauser and Revesz, 1989).

Finally, given that extending liability solely to site owners or solely to generators results in inefficiencies, liability regimes that target a larger set of parties can achieve two goals. First, such a regime can transmit proper incentives to a larger group of actors. Thus, even if a site owner should become insolvent, generators will have an incentive to continue monitoring. Second, dividing liability between a number of actors can decrease the likelihood of insolvency.

3.1.3. Cross-cutting issues

Three cross-cutting issues stand out in the normative analysis of environmental policy instrument choice: the implications of uncertainty; effects on technological change; and distributional considerations.

3.1.3.1. Implications of uncertainty for instrument choice The dual task facing policy makers of choosing environmental goals and selecting policy instruments to achieve those goals must be carried out in the presence of the significant uncertainty that affects the benefits and the costs of environmental protection. Since Weitzman's (1974) classic paper on "Prices vs. Quantities," it has been generally acknowledged that benefit uncertainty on its own has no effect on the identity of the efficient control instrument, but that cost uncertainty can have significant effects, depending upon the relative slopes of the marginal benefit (damage) and marginal cost functions. In particular, if uncertainty about marginal abatement costs is significant, and if marginal abatement costs are flat relative to marginal benefits, then a quantity instrument is more efficient than a price instrument.[102]

We rarely encounter situations in which there is exclusively either benefit uncertainty or cost uncertainty. On the contrary, in the environmental arena, we typically find that

[102] For an early empirical application in the environmental realm, see: Kolstad (1986).

the two are present simultaneously, and more often than not, it is benefit uncertainty that is of substantially greater magnitude. When marginal benefits are positively correlated with marginal costs (which, it turns out, is not uncommon), then there is an additional argument in favor of the relative efficiency of quantity instruments (Stavins, 1996). On the other hand, the regulation of stock pollutants will often favor price instruments, because the marginal benefit function—linked with the stock of pollution—will tend to be relatively flat, compared with the marginal cost function—linked with the flow of pollution (Newell and Pizer, 2003b).

In theory, there would be considerable efficiency advantages in the presence of uncertainty of hybrid systems—for example, quotas combined with taxes—or non-linear taxes[103] (Roberts and Spence, 1976; Weitzman, 1978; Kaplow and Shavell, 2002b; Pizer, 2002), but such systems have not been adopted.

3.1.3.2. Effects of instrument choice on technological change Environmental policy interventions foster constraints and incentives that affect the process of technological change (Kneese and Schulze, 1975; Orr, 1976). To be dynamically cost-effective, instruments need to foster rather than inhibit technological invention, innovation, and diffusion (Kemp and Soete, 1990). Both command-and-control and market-based instruments have the potential for forcing or inducing technological change, by requiring firms to alter their behavior. Technology and performance standards can be used to stimulate innovation by setting ambitious targets, beyond the reach of current technologies. But it is impossible to know whether a given target will be feasible or not, and so such policies run substantial risk of failure (Freeman and Haveman, 1972). Technology standards are particularly problematic, because they tend to freeze the development of technologies that might otherwise result in greater levels of control.

Much of the economic research on technological invention and innovation (commercialization) has focused on incentives for firm-level decisions to incur costs of research and development in the face of uncertain outcomes. The earliest relevant work was by Magat (1978, 1979), who compared taxes, subsidies, permits, effluent standards, and technology standards, and showed that all but technology standards would induce innovation biased toward emissions reduction. More recent theoretical attempts to rank policy instruments according to their innovation-stimulating effects (Fischer, Parry, and Pizer, 2003) conclude that an unambiguous rating of instruments is not possible. The ranking of instruments depends on the innovator's ability to appropriate spillover benefits of new technologies to other firms, the costs of innovation, environmental benefit functions, and the number of firms producing emissions (Carraro and Soubeyran, 1996; Katsoulacos and Xepapadeas, 1996; Montero, 2002).

Turning to technological diffusion (adoption), several theoretical studies have found that the incentive for the adoption of new technologies is greater under market-based

[103] In addition to the efficiency advantages of non-linear taxes, they also have the attribute of reducing the total (although not the marginal) tax burden of the regulated sector, relative to an ordinary linear tax, which is potentially important in a political economy context.

instruments than under direct regulation (Zerbe, 1970; Downing and White, 1986; Milliman and Prince, 1989), but theoretical comparisons among market-based instruments have produced only limited agreement (Milliman and Prince, 1989, 1992; Marin, 1991; Jung, Krutilla, and Boyd, 1996; Parry, 1998; Denicolò, 1999; Keohane, 1999). One overall result seems to be that auctioned permits are inferior in their diffusion incentives to emissions tax systems (but both are superior to command-and-control instruments).[104]

Closely related to the effects of instrument choice on technological change are the effects of vintage-differentiated regulation on the rate of capital turnover, and thereby on pollution abatement costs and environmental performance. Such vintage-differentiated regulation is a common feature of many environmental and other regulatory policies in the United States, wherein the standard for regulated units is fixed in terms of their date of entry, with later vintages facing more stringent regulation. In the most common application, commonly referred to as "grandfathering," units produced prior to a specific date are exempted from a new regulation or face less stringent requirements.

While this approach has long appealed to many participants in the policy community, economists have frequently noted that vintage-differentiated regulations can be expected—on the basis of standard investment theory—to retard turnover in the capital stock, and thereby to reduce the cost-effectiveness of regulation, compared with equivalent undifferentiated regulations. Furthermore, under some conditions the result can be higher levels of pollutant emissions than would occur in the absence of regulation. Such economic and environmental consequences are not only predictions from theory (Gruenspecht, 1981; Maloney and Brady, 1988); both types of consequences have been validated empirically in the context of specific regulations (Hartman, Bazdogan, and Nadkarni, 1979; Gruenspecht, 1982; Nelson, Tietenberg, and Donihue, 1993).

3.1.3.3. Distributional considerations Alternative policy instruments can have significantly different impacts on the distribution of benefits and costs. First with regard to benefits, taxes or tradeable permits can lead to localized "hot spots" with relatively high levels of ambient pollution. This is a significant distributional issue, and it can also become an efficiency issue if damages are non-linearly related to pollutant concentrations (Mendelsohn, 1986). The problem can be addressed, in theory, through the use of "ambient permits"[105] or through charge systems that are keyed to changes in ambient conditions at specified locations (Revesz, 1996), or through trading schemes that are simply constrained by the requirement that ambient standards not be violated.[106] Despite the theoretical literature on ambient systems going back to Montgomery (1972),

[104] For a detailed review of analyses of the effects of instrument choice on technological innovation and diffusion, see Jaffe, Newell, and Stavins (2002).

[105] Ambient permits entitle the owner to increase the concentration at a certain receptor site by a specified amount, rather than permitting some quantity of emissions.

[106] In theory, the locus of regulation can range from input levels (for example, through the use of a permit linked to the carbon content of fossil fuels), to emissions, to ambient concentrations, to exposure levels, to—ultimately—risk levels.

they have never been implemented,[107] with the partial exception of a two-zone trading system under Los Angeles' RECLAIM program.[108]

Turning to the cost side, taxes and auctioned tradeable permits can raise revenue for the government.[109] Revenue recycling (that is, using tax or permit revenues to reduce other, distortionary taxes) can significantly lower the costs of pollution control, relative to what the costs would be without such recycling (Jorgenson and Wilcoxen, 1994; Goulder, 1995b). It has been suggested by some that all of the abatement costs associated with a pollution tax can be eliminated through revenue recycling (Repetto et al., 1992), but environmental taxes can exacerbate distortions associated with remaining taxes on investment or labor, and research indicates that these distortions are at least as great as those from labor taxes (Bovenberg and de Mooji, 1994; Goulder, 1995a; Parry, 1995; Bovenberg and Goulder, 1996; Goulder, Perry, and Burtraw, 1997; Goulder et al., 1999).

Although distribution affects social welfare, an important strand of the theoretical literature suggests that distribution should not matter in choosing policy instruments. Hylland and Zeckhauser (1979) argue that given optimal taxation achieved by benefit-offsetting tax adjustments, maximizing net benefits should be the sole criterion for policy choice.[110] Despite its limitations, the income tax system is, in theory, best suited for redistributing income, with attempts at redistribution through other means causing inefficiencies that are at least as great as those encountered with income taxes. Moreover, Kaplow and Shavell (2001) demonstrate that any policy assessment that accords importance to non-utility criteria (including societal concerns for distribution) violates the Pareto principle, suggesting that environmental issues should be addressed ideally through a pair of policies: an efficient environmental policy instrument chosen solely on the basis of maximizing net benefits, and an income tax adjustment to offset possible undesirable distributional impacts.[111]

Political economy considerations may run counter to such theoretical arguments, since it is difficult to combine every environmental policy rule with a change in the income tax system. It may also be politically infeasible to adopt environmental policies that do not themselves address distributional concerns. Indeed, Arrow et al. (1996b) ar-

[107] Such systems can be difficult to implement. If there are many significant receptor sites, the implementation of tradeable permits will require separate markets for each type of permit. For a review of ambient permit approaches, see Tietenberg (1995).

[108] In the case of RECLAIM, empirical analysis indicated that a substantial share of the relevant heterogeneity in concentrations would be captured by employing just two zones (Johnson and Pekelney, 1996).

[109] While an allocation of permits made through sale or auction will have similar distributional consequences to a tax, a revenue-neutral emissions tax, in which revenues are refunded to regulated firms (but not in proportion to their emissions levels), can resemble—in distributional terms—a comparable tradeable permit program in which the permits are allocated without charge.

[110] Shavell (1981) and Kaplow and Shavell (1994) extended this result to legal rulemaking.

[111] See: Kaplow (1996, 2004). For a broader discussion of the underlying issues, see Kaplow and Shavell (2002a).

gue that the Kaldor–Hicks criterion should be considered as neither a necessary nor a sufficient condition for public policy.[112]

A policy's political feasibility is influenced strongly by its distributional implications. Auctioned permit systems or effluent charges can be more costly than comparable command-and-control instruments from the perspective of regulated firms. Tradeable permit systems, on the other hand, have the important attribute that in the absence of decreasing marginal transactions costs (essentially volume discounts), the equilibrium allocation and hence aggregate abatement costs of a tradeable permit system are independent of initial allocations (Stavins, 1995). Hence, the allocation decision can be left to politicians, with limited normative concerns about the potential effects of the chosen allocation on overall cost-effectiveness. In other words, cost-effectiveness or efficiency can be achieved, while distributional equity is simultaneously addressed with the same policy instrument.[113]

3.1.4. Normative lessons

Although there has been considerable experience in the United States with market-based instruments for environmental protection, this relatively new set of policy approaches has not replaced nor come anywhere close to replacing conventional, command-and-control policies. When and where these approaches have been used in their purest form and with some success, they have not always performed as anticipated. We review briefly the normative lessons that can be learned from research and experience.

3.1.4.1. Design and implementation The performance to date of market-based instruments for environmental protection provides evidence that these approaches can achieve major cost savings while accomplishing their environmental objectives. The performance of these systems also offers lessons about the importance of flexibility, simplicity, and the capabilities of the private sector.

In regard to flexibility, allowing flexible timing and intertemporal trading of permits—that is, banking allowances for future use—played a very important role in the SO_2 allowance trading program's performance (Ellerman et al., 1997), much as it did in the U.S. lead rights trading program a decade earlier (Kerr and Maré, 1997).[114] One of the most significant benefits of using market-based instruments may simply

[112] See the discussion of the Kaldor–Hicks criterion in section 2.1.1.

[113] This is one of the reasons why an international tradeable permit mechanism has been considered to be particularly attractive for addressing global climate change. Allocation mechanisms can be developed that address equity concerns of developing countries, and thus increase the political base for support, without jeopardizing the overall cost-effectiveness of the system. See, for example, Frankel (1999). It should be recognized, however, that in practice tradeable permits have typically not been allocated to achieve goals of distributional equity *per se*, but to achieve political feasibility (Joskow and Schmalensee, 1998).

[114] In theory, a fully cost-effective permit trading program must allow for both banking and borrowing (Rubin, 1996; Kling and Rubin, 1997).

be that technology standards are thereby avoided. Less flexible systems would not have led to the technological change that may have been induced by market-based instruments (Burtraw, 1996; Ellerman and Montero, 1998; Bohi and Burtraw, 1997; Keohane, 2001), nor the induced process innovations that have resulted (Doucet and Strauss, 1994).

In regard to simplicity, transparent formulae—whether for permit allocation or tax computation—are difficult to contest or manipulate. Requiring prior government approval of individual trades may increase uncertainty and transaction costs, thereby discouraging trading; these negative effects should be balanced against any anticipated benefits due to requiring prior government approval. Such requirements hampered EPA's Emissions Trading Program in the 1970s, while the lack of such requirements was an important factor in the success of lead trading (Hahn and Hester, 1989). In the case of SO_2 trading, the absence of requirements for prior approval reduced uncertainty for utilities and administrative costs for government, and contributed to low transactions costs (Rico, 1995).

One potentially important cause of the mixed performance of implemented market-based instruments is that many firms are simply not well equipped to make the decisions necessary to fully utilize these instruments. The focus of environmental, health, and safety departments in private firms has been primarily on problem avoidance and risk management, rather than on the creation of opportunities made possible by market-based instruments. Since market-based instruments have been used on a limited basis only, and firms are not certain that these instruments will be a lasting component on the regulatory landscape, it is not surprising that most companies have not reorganized their internal structure to fully exploit the cost savings these instruments offer (Reinhardt, 2000). Rather, most firms continue to have organizations that are experienced in minimizing the costs of complying with command-and-control regulations, not in making the strategic decisions allowed by market-based instruments.[115]

3.1.4.2. Identifying new applications Market-based policy instruments are considered today for nearly every environmental problem that is raised, ranging from endangered species preservation to global climate change.[116] Where the cost of abating pollution differs widely among sources, a market-based system is likely to have greater gains, relative to conventional, command-and-control regulations (Newell and Stavins, 2003). For example, it was clear early on that SO_2 abatement cost heterogeneity was great, because of differences in ages of plants and their proximity to sources of low-sulfur

[115] There are, of course, exceptions. See: Hockenstein, Stavins, and Whitehead (1997). There is anecdotal evidence which may suggest that the existence of tradeable permit programs is changing the way firms evaluate environmental risk (Hartridge, 2003; Tietenberg, 2003).

[116] See, for example, Goldstein (1991) on species protection, and Fisher et al. (1996); Hahn and Stavins (1995); Schmalensee (1998); and Stavins (1997) on applications to global climate change. More broadly, see: Ayres (2000).

coal. But where abatement costs are more uniform across sources, the political costs of enacting an allowance trading approach are less likely to be justifiable.

The choice of market-based instrument should depend on the characteristics of the pollutant, the degree of uncertainty, expected changes in the economic environment, ability to induce technological change, potential transactions costs, and other significant interacting factors. Finally, considerations of political feasibility point to the wisdom (more likely success) of proposing market-based instruments when they can be used to facilitate a cost-effective, aggregate emissions reduction (as in the case of the SO_2 allowance trading program in 1990), as opposed to a cost-effective reallocation of the status quo burden.

3.2. Positive issues and analysis

A set of positive political economy questions are raised by the increasing use of market-based instruments for environmental protection.[117] First, why was there so little use of market-based instruments in the United States, relative to command-and-control instruments, over the 30-year period of major environmental regulation that began in 1970, despite the apparent advantages these instruments offer? Second, when market-based instruments have been adopted, why has there been such great reliance on tradeable permits allocated without charge, despite the availability of a much broader set of incentive-based instruments? Third, why has the political attention given to market-based environmental policy instruments increased dramatically in recent years?

To examine these questions, we employ a "political market" metaphor (Keohane, Revesz, and Stavins, 1998). The commodity being supplied is legislators' support for a given policy instrument, and the currency is resources that can be used for re-election—contributions, endorsements, votes, and other forms of support. Demand for legislative outcomes comes from interest groups, including environmental advocacy organizations, private firms, industry associations, organized labor, and consumers. Ultimately, the choice of environmental policy instrument is determined by the equilibrium between demand by interest groups and supply by legislators and regulators.

3.2.1. Historical dominance of command-and-control

On the regulatory demand side, affected firms and their trade associations prefer instruments that have lower aggregate costs for their industry, or that increase aggregate profits by creating rents or barriers to entry. An individual firm may actually prefer regulation to the status quo if that regulation gives the firm a competitive advantage over rivals.[118] Command-and-control standards have the potential to

[117] This section of the chapter draws upon: Keohane, Revesz, and Stavins (1998); Hahn and Stavins (1991); and Stavins (1998).

[118] There are other possible explanations for firms' preferences, including the possibility that existing agents tend to support the status quo for fear that their expertise will be devalued under new regimes. There is also

generate rents for existing firms in an industry, increasing aggregate profits. Regulations that establish long-term barriers to entry can sustain these profits indefinitely and will be strongly preferred by industry groups (Buchanan and Tullock, 1975; Maloney and McCormick, 1982). Command-and-control standards are inevitably set up with extensive input from industry, and frequently contain more stringent requirements for new sources and other effective barriers to entry (Stigler, 1971; Rasmusen and Zupan, 1991). Firms also tend to favor command-and-control regulation or grandfathered permits over pollution taxes or auctioned permits because they shift less of the distributional burden onto private industry (Arnold, 1995; Crandall, 1983; Hahn and Noll, 1990). Auctioned permits and pollution taxes require firms to pay not only abatement costs, but also regulatory costs in the form of permit purchases or tax payments.

Environmental advocacy groups are also on the demand side of the market for legislative support. Such groups seek to maximize their utility, which depends on both their organizational well-being and the level of environmental quality. For a long time, nearly all environmental advocacy groups were actively hostile towards market-based instruments. One reason was philosophical: environmentalists frequently perceived pollution taxes and tradeable permits as "licenses to pollute." Although such ethical objections to the use of market-based environmental strategies have greatly diminished, they have not disappeared completely (Sandel, 1997).[119] A second concern was that damages from pollution were difficult or impossible to quantify and monetize, and thus could not be summed up in a marginal damage function or captured by a Pigovian tax rate (Kelman, 1981a). Third, environmental organizations have opposed market-based schemes for strategic reasons, particularly the fear that permit levels and tax rates—once implemented—would be more difficult to tighten over time than command-and-control standards. For example, if permits are given the status of "property rights," then any subsequent attempt by government to reduce pollution levels further could meet with demands for compensation.[120] Finally, environmental organizations have objected to decentralized instruments on the technical grounds that even if emission taxes or tradeable permits reduce overall levels of emissions, they can—in theory—lead to localized "hot spots" with relatively high levels of ambient pollution.

The final influential group demanding support from legislators is organized labor. Labor groups can be expected to seek protection for jobs, and they may oppose instruments that are likely to lead to plant closings or major industrial dislocations. For

the possibility that market-based instruments were opposed simply because they were not well understood (Kelman, 1981b).

[119] Sandel (1997) argues that emissions trading will foster "immoral" behavior by giving firms a "license to pollute," despite the fact that pollution taxes and tradeable permits create incentives for firms to decrease pollution. See Shavell et al. (1997) for replies to Sandel's arguments. For a broader examination of the ethical limitations of markets in other arenas, see Sandel (1998).

[120] This concern was alleviated in the SO_2 provisions of the Clean Air Act Amendments of 1990 by an explicit statutory provision that permits do not represent property rights.

example, labor might oppose a tradeable permit scheme in which firms would tend to close factories in heavily polluted areas, sell permits, and relocate to less polluted areas where permits are cheaper (Hahn and Noll, 1990). In the case of restrictions on clean air, organized labor has taken the side of the United Mine Workers, whose members are heavily concentrated in eastern mines that produce higher-sulfur coal, and had therefore opposed pollution-control measures that would increase incentives for using low-sulfur coal from the largely non-unionized (and less labor-intensive) mines in Wyoming's and Montana's Powder River Basin. In the 1977 debates over amendments to the Clean Air Act, organized labor fought to include a command-and-control standard that effectively required scrubbing, thereby seeking to discourage switching to cleaner western coal (Ackerman and Hassler, 1981). Likewise, the United Mine Workers opposed the SO_2 allowance trading system in 1990, because of a fear that it would encourage a shift to western low-sulfur coal from non-unionized mines.

Turning to the supply side of environmental regulation, legislators may be thought of as providing support as a function of the opportunity cost of supporting a given instrument, the psychological cost associated with their ideological preferences, and the losses or gains of constituency support as a result of an action (Keohane, Revesz, and Stavins, 1998). Legislators have had a number of reasons to find command-and-control standards attractive. First, many legislators and their staffs are trained in law, which may predispose them to favor conventional regulatory approaches and lead to a status quo bias in favor of command-and-control approaches (Kneese and Schulze, 1975). Second, standards tend to help hide the costs of pollution control (McCubbins and Sullivan, 1984; Hahn, 1987), while market-based instruments generally impose those costs more directly. Third, standards offer greater opportunities for symbolic politics, because strict standards—strong statements of support for environmental protection—can readily be combined with less visible exemptions or with lax enforcement measures.

Fourth, if politicians are risk averse, they will prefer instruments that involve more certain effects.[121] The flexibility inherent in market-based instruments creates uncertainty about distributional impacts. Typically, legislators in a representative democracy are more concerned with the geographic distribution of costs and benefits than with comparisons of total benefits and costs. Hence, aggregate cost-effectiveness—the major advantage of market-based instruments—is less likely to play a significant role in the legislative calculus than whether a politician is getting a good deal for his or her constituents (Shepsle and Weingast, 1984).

Finally, legislators are wary of enacting programs that are likely to be undermined by bureaucrats in their implementation. And bureaucrats are less likely to undermine legislative decisions if their own preferences over policy instruments are accommodated. Bureaucratic preferences—at least in the past—were not supportive of market-based

[121] Legislators are likely to behave as if they are risk averse, even if they are personally risk neutral, if their constituents punish unpredictable policy choices or their reelection probability is nearly unity (McCubbins, Noll, and Weingast, 1989).

instruments. Government bureaucrats—like their counterparts in environmental advocacy groups and trade associations—opposed market-based instruments to prevent their expertise from becoming obsolete, that is, to preserve their human capital (Hahn and Stavins, 1991).[122]

3.2.2. Prevalence of tradeable permits allocated without charge

Economic theory suggests that the choice among tradeable permits, pollution taxes, and other market-based instruments should be based upon case-specific factors, but major applications in the United States have nearly always taken the form of tradeable permits allocated without charge, rather than through auctions,[123] despite the apparent economic superiority of the latter mechanism in terms of economic efficiency. Many participants in the policy process have reasons to favor tradeable permits allocated without charge over other market-based instruments.

On the regulatory demand side, existing firms favor tradeable permits allocated without charge because such permits convey rents to firms. Moreover, like stringent command-and-control standards for new sources, but unlike auctioned permits or taxes, permits allocated without charge give rise to entry barriers, since new entrants must purchase permits from existing holders. Thus, the rents conveyed to the private sector by tradeable permits allocated without charge are, in effect, sustainable.

Environmental advocacy groups have generally supported command-and-control approaches, but given the choice between tradeable permits and emission taxes, these groups strongly prefer the former. Environmental advocates have a strong incentive to avoid policy instruments that make the costs of environmental protection highly visible to consumers and voters; and taxes make those costs more explicit than permits. Also, environmental advocates prefer permit schemes because they specify the quantity of pollution reduction that will be achieved, in contrast with the indirect effect of pollution taxes. Overall, some environmental groups have come to endorse the tradeable permits approach because it promises the cost savings of pollution taxes, but without the drawbacks that environmentalists associate with tax instruments.

Tradeable permits allocated without charge are easier for legislators to supply than taxes or auctioned permits, again because the costs imposed on industry are less visible and less burdensome, since no money is exchanged at the time of the initial permit allocation. Also, permits allocated without charge offer a much greater degree of political control over the distributional effects of regulation, facilitating the formation of majority coalitions. Joskow and Schmalensee (1998) examined the political process of

[122] Subsequently, this same incentive led EPA staff involved in the acid rain program to become strong proponents of trading for a variety of other pollution problems.

[123] The EPA does have an annual (revenue-neutral) auction of SO_2 allowances, but this represents less than 2 percent of the total allocation (Bailey, 1996). While the EPA auctions may have helped in establishing the market for SO_2 allowances, they are a trivial part of the overall program (Joskow, Schmalensee, and Bailey, 1998).

allocating SO_2 allowances in the 1990 amendments, and found that allocating permits on the basis of prior emissions can produce fairly clear winners and losers among firms and states. An auction allows no such political maneuvering.

3.2.3. Increased attention to market-based instruments

Interest in and use of incentive-based instruments has increased at both the Federal and state levels in recent years (Hahn, 2000). Given the historical lack of receptiveness by the political process to market-based approaches to environmental protection, why has there been this rise in the use of these approaches? It would be gratifying to believe that increased understanding of market-based instruments had played a large part in fostering their increased political acceptance, but how important has this really been?

In 1981, Kelman surveyed Congressional staff members, and found that support and opposition to market-based environmental policy instruments was based largely on ideological grounds: Republicans, who supported the concept of economic-incentive approaches, offered as a reason the assertion that "the free market works," or "less government intervention" is desirable, without any real awareness or understanding of the economic arguments for market-based programs. Likewise, Democratic opposition was largely based upon ideological factors, with little or no apparent understanding of the real advantages or disadvantages of the various instruments (Kelman, 1981b). What would happen if we were to replicate Kelman's survey today? Our refutable hypothesis is that we would find increased support from Republicans, greatly increased support from Democrats, but insufficient improvements in understanding to explain these changes.[124] So what else has mattered?

First, one factor has surely been increased pollution control costs, which have led to greater interest from all parties in cost-effective instruments. By the late 1980's, even political liberals and environmentalists were beginning to question whether conventional regulations could produce further gains in environmental quality. During the previous twenty years, pollution abatement costs had continually increased, as stricter standards moved the private sector up the marginal abatement-cost function. By 1990, U.S. pollution control costs had reached \$125 billion annually, nearly a 300% increase in real terms from 1972 levels (U.S. Environmental Protection Agency, 1990; Jaffe et al., 1995). Market-based instruments represent an effective way to reduce aggregate abatement costs.

Second, a factor that became important in the late 1980's was strong and vocal support from some segments of the environmental community. By supporting tradeable permits for acid rain control, the Environmental Defense Fund (EDF) seized a market niche in the environmental movement, and successfully distinguished itself from

[124] But there has been some increased understanding of market-based approaches among policy makers. This has partly been due to increased understanding by their staffs, a function—to some degree—of the economics training that is now common in law schools, and of the proliferation of schools of public policy (Hahn and Stavins, 1991).

other groups.[125] Related to this, a third factor was that the SO_2 allowance trading program, the leaded gasoline phasedown, and the CFC phaseout were all designed to *reduce* emissions, not simply to *reallocate* them cost-effectively among sources. Market-based instruments are most likely to be politically acceptable when proposed to achieve environmental improvements that would not otherwise be feasible (politically or economically).

Fourth, deliberations regarding the SO_2 allowance system, the lead system, and CFC trading differed from previous attempts by economists to influence environmental policy in an important way: the separation of ends from means, i.e. the separation of consideration of goals and targets from the policy instruments used to achieve those targets. By accepting the politically identified (and potentially inefficient) goal, the ten-million ton reduction of SO_2 emissions, for example, economists were able to focus successfully on the importance of adopting a cost-effective means of achieving that goal. Fifth, acid rain was an unregulated problem until the SO_2 allowance trading program of 1990; and the same can be said for leaded gasoline and CFC's. Hence, there were no existing constituencies—in the private sector, the environmental advocacy community, or government—for the *status quo* approach, because there was no *status quo* approach.

Sixth, by the late 1980's, there had already been a perceptible shift of the political center toward a more favorable view of using markets to solve social problems. The George H. W. Bush Administration, which proposed the SO_2 allowance trading program and then championed it through an initially resistant Democratic Congress, was (at least in its first two years) "moderate Republican;" and phrases such as "fiscally responsible environmental protection" and "harnessing market forces to protect the environment" do have the sound of quintessential moderate Republican issues.[126] But, beyond this, support for market-oriented solutions to various social problems had been increasing across the political spectrum for the previous fifteen years, as was evidenced by deliberations on deregulation of the airline, telecommunications, trucking, railroad, and banking industries. Indeed, by the mid-1990s, the concept (or at least the phrase), "market-based environmental policy," had evolved from being politically problematic to politically attractive.

[125] When the memberships (and financial resources) of other environmental advocacy groups subsequently declined with the election of the environmentally-friendly Clinton-Gore Administration, EDF continued to prosper and grow (Lowry, 1993). EDF has since renamed itself "Environmental Defense."

[126] The Reagan Administration enthusiastically embraced a market-oriented ideology, but demonstrated little interest in employing actual market-based policies in the environmental area. From the Bush Administration through the Clinton Administration, interest and activity regarding market-based instruments—particularly tradeable permit systems—continued to increase, although the pace of activity in terms of newly implemented programs declined during the Clinton years, when a considerable part of the related focus was on global climate policy (Hahn, Olmstead, and Stavins, 2003).

4. Allocation of responsibility across levels of government

Throughout most of U.S. history, state and local governments have had primary responsibility for health-and-safety regulation, including environmental protection,[127] but since 1970, the Federal government has played an increasingly important role in environmental regulation (Revesz, 2001a). What regulatory advantages or disadvantages does the Federal government have, compared with state governments? What does this suggest about how regulatory responsibility should be allocated?

4.1. Positive review of responsibility of levels of government

Before 1970, Congress largely left environmental regulation to the states. As the modern environmental movement gained political force, however, the Federal government began assembling its regulatory framework. Congress's first major effort came in 1969 with the passage of the National Environmental Policy Act,[128] which laid out broad environmental goals and required Federal agencies to assess the environmental impacts of their programmatic actions.

A set of laws passed over the subsequent two decades marked the federal government's new-found commitment to environmental regulation. Three statutes formed the backbone of the federal scheme: the Clean Air Act of 1970,[129] the Clean Water Act of 1972,[130] and the Resource Conservation and Recovery Act of 1976,[131] all of which have been amended numerous times since their adoption. These laws, characterized by their command-and-control approaches to regulation, granted wide discretion to the newly-created Environmental Protection Agency to set tolerance levels for various pollutants. In addition, Congress protected endangered species,[132]set limits on contaminants allowed in drinking water,[133]and created a system of strict joint-and-several liability for parties responsible for abandoned hazardous waste sites.[134]

Federal environmental laws typically (but with important exceptions) establish minimum environmental standards while leaving states free to adopt more stringent standards. Many states have done exactly that. Some have adopted tighter thresholds for automobile emissions; others have created their own "Superfund" programs; and others have implemented their own state-based environmental protection acts (Revesz, 2001a). States also remain free to regulate in areas the federal government has opted not to, such as wetlands preservation or groundwater quality.

[127] See Hillsborough County v. Automated Medical Lab., Inc., 471 U.S. 707, 715–16 (1985).

[128] 42 U.S.C. §§4321–4370a.

[129] 42 U.S.C. §§7401–7642.

[130] 33 U.S.C. §§1251–1376. The Act was originally titled the Federal Water Pollution Control Act.

[131] 42 U.S.C. §§6901–6987.

[132] Endangered Species Act, 16 U.S.C. 1531–1544.

[133] Safe Drinking Water Act, 42 U.S.C. 300f *et seq*.

[134] Comprehensive Environmental Response, Compensation, and Liability Act of 1980, 42 U.S.C. §9601 *et seq*.

4.2. Normative review of allocation of regulatory responsibility

A rebuttable presumption in favor of decentralized authority over environmental regulation may be posited for three reasons. First, in a large and diverse country, different regions will likely have different environmental preferences. Second, the benefits of environmental protection vary throughout the country. For example, a stringent air quality standard may benefit many people in densely populated areas but only a few elsewhere. Third, the costs of meeting a given standard differ across geographic regions.

Federal intervention in environmental regulation has traditionally been justified by reference to one or more of three perceived pathologies that hamper effective state regulation: the race to the bottom induced by competition for mobile resources; the existence of significant interstate externalities; and the public-choice rationale that environmental groups can more effectively lobby at the Federal level than at the state level. Analysis has cast doubt on the viability of these justifications.[135]

4.2.1. Competition among political jurisdictions: the race to the bottom

The conventional race-to-the-bottom rationale for federal regulation posits that states, in an effort to induce geographically mobile firms to locate within their jurisdictions, will offer them sub-optimally lax environmental standards in order to benefit from additional jobs and tax revenues. In the absence of Federal regulation, states would therefore systematically under-regulate.

4.2.1.1. Normative assessment of the race-to-the-bottom claim The theoretical foundation for the view that interstate competition for industry would inevitably lead to sub-optimally lax environmental standards is weak. Indeed, economic analysis of the effects of interstate competition on the choice of environmental standards indicates that rather than a race to the bottom, inter-jurisdictional competition may be expected to lead to the maximization of social welfare, at least under conditions of perfect competition (Oates and Schwab, 1988). In their model, Oates and Schwab posit jurisdictions that compete for mobile capital through the choice of taxes and environmental standards. A higher capital stock benefits residents in the form of higher wages, but hurts them through foregone tax revenues and lower environmental quality. Each jurisdiction makes two policy decisions: it sets a tax rate on capital and an environmental standard. Oates and Schwab show that competitive jurisdictions will set a net tax rate on capital of zero (the rate that exactly covers the cost of public services provided to the capital, such as police and fire protection). In turn, competitive jurisdictions will set an environmental standard that is defined by equating the willingness to pay for an additional unit of environmental quality with the corresponding change in wages. Oates and Schwab show that these choices of tax rates and environmental standards are efficient.

[135] This section draws upon Revesz (2001b). Also see Krier (1995).

When the assumption of perfect competition is relaxed, strategic interactions among the states can lead to under-regulation absent federal intervention. But it is likewise plausible that in other instances the reverse would be true: that the strategic interactions among the states would lead to over-regulation absent federal intervention. Accordingly, there is no compelling race-to-the-bottom justification for across-the-board federal minimum standards, the cornerstone of Federal environmental law.

The most extensive analyses of the effects of imperfect competition among states show that either over-regulation or under-regulation can result (Markusen, Morey, and Olewiler, 1993, 1995), depending on the levels of firm-specific costs, plant-specific costs, and transportation costs. Similarly, if a firm has market power enabling it to affect prices, it will be able to extract a sub-optimally lax standard; but if a state has market power, the reverse would be true (Revesz, 1992, 1997b). In summary, just as there are situations in which interstate competition produces environmental under-regulation (Esty and Geradin, 2001), there are other plausible scenarios under which the result is over-regulation.

Moreover, even if states systematically enacted sub-optimally lax environmental standards, Federal environmental regulation would not necessarily improve the situation. If states cannot compete over environmental regulation because it has been federalized, they will compete along other regulatory dimensions, leading to sub-optimally lax standards in other areas, or along the fiscal dimension, leading to the under-provision of other public goods. Thus, the reduction in social welfare implicit in race-to-the-bottom arguments would not be eliminated by federalizing environmental regulation. Rather, the federalization of all regulatory and fiscal decisions would be necessary to solve the problem.[136]

Several authors have attempted to rehabilitate some version of the race-to-the-bottom justification for Federal regulation. Their arguments, however, rely on conflations of alternative justifications for environmental regulation, such as the presence of interjurisdictional externalities or public choice failures (Esty, 1996; Esty and Geradin, 2001), unsupported public-choice rationales that are analytically distinct from the race-to-the-bottom justification (Swire, 1996), weak empirical support (Engel, 1997), or circular notions that Federal environmental regulation serves to reinforce “national evaluative norms” (Sarnoff, 1997). The critics therefore fail to address two core difficulties confronting supporters of Federal environmental regulation (Revesz, 1997b). First, none are able to explain why Federal environmental floors are an appropriate response to races that can lead *either* to over-regulation or under-regulation. Regulatory ceilings would be, after all, the appropriate Federal response to widespread over-regulation. Second, their arguments for federalizing environmental decision-making prove too much, and tend equally to support the claim that all state fiscal and regulatory decisions should be addressed at the Federal level.

[136] Similarly, there is a concern that absent federal regulation, firms could capture rents created by locational advantages that otherwise would accrue to the states. But if environmental regulation is federalized, the rents could be captured with respect to another component of costs. Only complete centralization would address the problem (Engel and Rose-Ackerman, 2001).

4.2.1.2. Positive assessment of the race-to-the-bottom claim The validity of race-to-the-bottom arguments for federal regulation cannot be resolved on theoretical grounds alone. Empirical analysis is required, and available evidence indicates that the stringency of environmental regulation does not have a statistically significant effect on plant location decisions (Bartik, 1988b, 1989; Friedman, Gerlowski, and Silberman, 1992; Levinson, 1996; McConnell and Schwab, 1991).

More generally, the empirical economic literature on the effects of environmental regulation provides little evidence to support the hypothesis that environmental regulation has a significant adverse effect on economic growth or on other measures of competitiveness (Jaffe et al., 1995). Recent studies have reinforced this conclusion, finding that environmental regulation does not reduce labor demand (Berman and Bui, 2001a), and does not impair productivity (Berman and Bui, 2001b). Such findings are not surprising, given that for all but the most heavily polluting industries, the costs of complying with environmental regulation are a small share of the total costs of production—an average of about 2 percent (Jaffe et al., 1995). It follows that the difference in production costs among jurisdictions with relatively more stringent and relatively less stringent environmental standards is even less. Other regulatory factors—including the level of state taxes, the provision of public services, and the degree of unionization of a state's labor force—have been shown empirically to exert significant influences on location decisions, just as environmental regulations have not (Bartik, 1988b, 1989; Levinson, 1996). Moreover, there is evidence that large national or multinational firms build their plants to meet the standards of the most stringent jurisdiction in which they have production facilities. Thus, they do not benefit from lower costs of environmental regulation when they operate in jurisdictions with laxer standards (Jaffe et al., 1995).

Of course, even if empirical evidence indicated that firms move from or do not locate in jurisdictions with more stringent environmental standards, this would not necessarily indicate that such a "race-to-the-bottom" was welfare-decreasing. A study of firm mobility measures only what states lose as a result of more stringent environmental standards; it does not assess the corresponding gains that may result from better environmental quality. A state that makes its environmental standards more stringent and thereby loses some economic activity may well increase its social welfare, if the environmental gains are greater than the losses.

4.2.2. Transboundary environmental problems

In contrast to the race-to-the-bottom argument, the presence of interstate externalities provides a potentially sound theoretical argument for Federal regulation. A state that sends pollution to another state can obtain the benefits of the economic activity that generates the pollution, but not suffer the full costs of that activity (Revesz, 1996). Transaction costs—particularly in the case of air pollution—are likely to be sufficiently high to prevent the formation of interstate compacts.[137]

[137] It is difficult for such compacts to emerge in the absence of a clearly defined baseline regarding when upwind states have the right to send pollution downwind, and in the absence of generally accepted models

The fact that interstate externalities provide a sensible justification for federal intervention does not mean that existing federal environmental regulations can be justified on these grounds. For some environmental problems, such as the control of drinking water quality, there are virtually no interstate externalities; the effects are almost exclusively local. Even with respect to problems for which interstate externalities exist, the rationale calls only for a response well targeted to the problem, such as a limit on the quantity of pollution that can cross state lines, rather than across-the-board Federal regulation.

In fact, the environmental statutes have been an ineffective response to the problem of interstate externalities. The core of the Clean Air Act, which addresses the type of pollution for which externalities are believed to be most prevalent, consists of a series of Federally prescribed ambient standards and emissions standards.[138] The federal emission standards do not effectively combat the problem of interstate externalities, because they do not regulate the number of sources within a state or the location of those sources. Similarly, the federal ambient air quality standards are not well targeted to address the problem of interstate externalities, since they require states to restrict pollution that may have only in-state consequences, and states can meet the ambient standards but still export pollution to downwind states (through tall stacks or locations near the interstate border). In fact, a state might meet its ambient standards precisely because it exports a large proportion of its pollution.[139] Sections 110(a)(2)(D) and 126(b),[140] enacted in 1977, are the only provisions of the Clean Air Act specifically designed to combat interstate externalities. They create a mechanism by which downwind states can seek to enjoin excessive upwind pollution. During the first two decades of the program, however, no downwind state prevailed on such a claim.

4.2.3. Public choice and systematic bias

Advocates for Federal regulation on public choice grounds typically assert that state political processes undervalue the benefits of environmental regulation, or overvalue the corresponding costs. Even if this is true, of course, it does not follow that federalizing environmental law will necessarily provide a solution. Federal regulation is justifiable only if the outcome at the Federal level is socially more desirable, either because there

for translating a source’s emissions into ambient air quality degradation. Moreover, for different pollution sources, the range of affected states will vary, rendering less likely the emergence of conditions favoring cooperation.

[138] 42 U.S.C. §§7409, 7411 (1994).

[139] The Federal environmental statutes have exacerbated, rather than ameliorated, the problem of interstate externalities. In the context of the Clear Air Act, the Federal ambient standards give states an incentive to encourage sources within their jurisdiction to use taller stacks (or to locate close to downwind borders). Not surprisingly, the use of tall stacks expanded considerably after the passage of the Clean Air Act in 1970, when only two stacks in the United States were higher than 500 feet. By 1985, more than 180 stacks were higher than 500 feet, and 23 were higher than 1,000 feet (Revesz, 1996).

[140] 42 U.S.C. §§7410(D), 7426 (1994).

is less under-regulation or because any over-regulation leads to smaller social welfare losses. There are several reasons for being skeptical about the soundness of such a claim.

4.2.3.1. Normative foundation for public choice claims The public choice mechanism that makes it possible for Federal regulation to correct for under-regulation at the state level is far from self-evident. For example, Esty (1996) states that "[a]t the centralized level, environmental groups find it easier to reach critical mass and thereby to compete on more equal footing with industrial interests." He adds that the difficulty of mobilizing the public in many separate jurisdictions is well established. In fact, the logic of collective action may suggest the opposite: given the costs of organizing necessarily larger groups at the Federal level, those groups will likely prove less effective there than at the state level. Aggregating environmental interests on a national level increases the heterogeneity of environmental policy priorities, thereby complicating organizational challenges. The situation is likely to be different for regulated industry groups, which frequently consist of firms with nationwide operations. For such firms, operating at the Federal level poses no additional free-rider problems or loss of homogeneity.

The relevant question is whether the additional problems faced by environmental groups at the Federal level are outweighed by benefits arising from the fact that the clash of interest groups takes place before a single legislature, a single administrative agency, and, in part, as a result of the exclusive venue of the U.S. Court of Appeals for the D.C. Circuit over important environmental statutes, in a single court (Revesz, 1997a). It might be the case that economies of scale of operating at the Federal level would more than outweigh the increased free-rider problems. Such economies of scale might well hold for certain costs associated with effective participation in the regulatory process, such as for hiring a competent scientist. But the structure of other political costs is likely to be quite different. For example, with respect to access to the legislative process, a standard public choice account is that the highest bidder prevails (Peltzman, 1976; Stigler, 1971). Thus, the benefit that a party receives from its expenditures is a function of the expenditures of the other party. Unless costs of this type are small, economies of scale of operating at the federal level are unlikely to outweigh the additional free-rider problems.

Given the standard public choice argument for federal environmental regulation, it is not clear why the problems observed at the state level would not be replicated at the Federal level (Revesz, 1997c). The logic of collective action would suggest that the large number of citizen-breathers, each with a relatively small stake in the outcome of a particular standard-setting proceeding, will be overwhelmed in the political process by concentrated industrial interests with a large stake in the outcome. This problem could occur at the Federal level as well as at the state level.

4.2.3.2. Positive support for public choice pathologies Public choice arguments for federal regulation rest on two empirical claims concerning the nature of state regulatory actions: (1) that states ignored environmental problems before 1970, when the major environmental statutes began to be enacted; and (2) that states continue to be less concerned about environmental problems than is the Federal government.

First, the view that the states ignored environmental problems before 1970 is simply not correct. Several studies show that during the 1960s, without Federal prodding, states were making considerable strides with respect to the control of air pollution. In particular, the concentrations of important air pollutants were falling at significant rates, and the number of states, counties, and municipalities with regulatory programs to control air pollution was increasing rapidly (Crandall, 1983; Goklany, 1998a, 1998b; Portney, 1990; Stern, 1982). For example, sulfur dioxide concentrations fell by about 11 percent annually between 1964 and 1971, but only by about 5 percent per year in the decade after the Federal government began regulating. Similarly, concentrations of total suspended particulates dropped sharply during the 1960s, but the pace of reduction slowed significantly in the 1970s (Crandall, 1983). The genesis of Federal environmental regulation is consistent with this evidence. The Clean Air Act of 1970 was a response to industry pressure for Federal regulation as a means of discouraging states from setting more stringent (and hence non-uniform) standards (Elliott, Ackerman, and Millian, 1985).

The second view—that states are less concerned about environmental problems than the Federal government—can be countered with reference to many states' innovative environmental laws which impose (sometimes significant) constraints on in-state firms. In the areas of automobile emission standards, hazardous waste regulation, non-hazardous solid waste regulation, and wetlands protection, states have taken an active role in (effectively) regulating to improve environmental quality, and this involvement increased in the 1990s in the face of the Federal government's less aggressive action on environmental matters.

Clearly not every state is equally active in the environmental regulatory arena. The citizens of some states may have preferences for laxer environmental regulation than the Federal regulatory level and may therefore not have any reason to adopt voluntarily additional environmental constraints. Indeed, an analysis of Federal representatives' voting records on issues of environmental concern indicates a strong correlation between support for "pro-environment" bills in Congress and heightened in-state environmental regulatory programs. The existence of significant state regulation calls into question the simplistic public choice claim that environmental groups are less able to lobby effectively at the state level than at the federal level (Revesz, 2001a).

5. Conclusions

The growing use of economic analysis to inform environmental decision making marks increasing acceptance of the usefulness of these tools to help focus and improve regulation. Debates about the normative standing of the Kaldor–Hicks criterion and the challenges inherent in making benefit–cost analysis operational will likely continue.

Nevertheless, economic analysis has assumed a significant position in the regulatory state.

At the same time, despite the arguments made for decades by economists and others, there seems to be no more than limited political support in the United States for much broader use of benefit–cost analysis to assess proposed or existing environmental regulations. In truth, these analytical methods remain on the periphery of policy formulation. In fact, as long as leaders on both sides of the debates in the policy community continue to react on ideological bases to proposals for such "regulatory reform," the status quo is unlikely to change. Perhaps the significant changes that have taken place over the past twenty years with regard to the means of environmental policy—that is, acceptance of market-based environmental instruments—can provide a model for progress with regards to analysis of the ends—the targets and goals—of public policies in this domain.

Certainly the change has been dramatic. Market-based instruments have moved center stage, and policy debates today look very different from those twenty years ago, when these ideas were routinely characterized as "licenses to pollute" or dismissed as completely impractical. Market-based instruments are now considered seriously for each and every environmental problem that is tackled, ranging from endangered species preservation to regional smog to global climate change. It is reasonable to anticipate that market-based instruments will enjoy increasing acceptance in the years ahead.

Of course, no particular form of government intervention, no individual policy instrument—whether market-based or conventional—and no specific level of government is appropriate for all environmental problems. Which instrument or level of government is best in any given situation depends upon a variety of characteristics of the environmental problem, and the social, political, and economic context in which it is being regulated. There is no policy panacea. But economic instruments are now part of the available policy portfolio, and ultimately that is good news both for environmental protection and economic well-being.

Acknowledgements

Katharine Emans, Nicholas Bagley, and Gernot Wagner provided valuable research assistance, and helpful comments on previous versions were provided by Dallas Burtraw, Daniel Esty, Myrick Freeman, Robert Hahn, Louis Kaplow, Nathaniel Keohane, Charles Kolstad, Peter Menell, Richard Newell, Sheila Olmstead, Mitchell Polinsky, Paul Portney, Steven Shavell, and Kerry Smith. The authors alone are responsible for any remaining errors.

References

Abraham, K.L. (1988). "Environmental Liability and the Limits of Insurance". Columbia Law Review 88, 942–988.

Ackerman, B.A., Hassler, W.T. (1981). Clean Coal/Dirty Air. Yale University Press, New Haven, CT.

Ackerman, F., Heinzerling, L. (2002). "Pricing the Priceless: Cost-Benefit Analysis of Environmental Protection". University of Pennsylvania Law Review 150, 1553–1584.

Adler, M.D., Posner, E.A. (Eds.) (2001). Cost-Benefit Analysis: Legal, Economic and Philosophical Perspectives. University of Chicago Press, Chicago, IL.

Ainslie, G. (1991). "Derivation of Rational Economic Behavior from Hyperbolic Discount Curves". American Economic Review 81, 334–340.

Alberini, A., Cropper, M., Krupnick, A., Simon, N.B. (2004). "Does the value of a statistical life vary with age and health status? Evidence from the US and Canada". Journal of Environmental Economics and Management 48 (1), 769–792.

Ananathanarayanan, A. (1998). "Is There a Green Link? A Panel Data Value Event Study of the Relationship Between Capital Markets and Toxic Releases", Working paper, Rutgers University.

Anderson, R.C., Hofmann, L.A., Rusin, M. (1990). "The Use of Economic Incentive Mechanisms in Environmental Management", Research Paper 51, American Petroleum Institute, Washington, D.C.

Anderson, S.T., Newell, R.G. (2004). "Information Programs for Technology Adoption: The Case of Energy-Efficiency Audits". Resource and Energy Economics 26 (1), 27–50.

Anderson, T.L. (1983). Water Crisis: Ending the Policy Drought. Cato Institute, Washington, D.C.

Anderson, T.L., Leal, D.R. (1991). Free Market Environmentalism. Westview Press, Boulder, CO.

Andreoni, J. (1995). "Warm-Glow versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments". Quarterly Journal of Economics 110 (1), 1–21.

Ariansen, P. (1998). "Anthropocentrism with a human face". Ecological Economics 24 (2–3), 153–162.

Arnold, F.S. (1995). Economic Analysis of Environmental Policy and Regulation. Wiley, John and Sons, Inc., New York.

Arrow, K. (1963). Social Choice and Individual Values, 2nd edn. Yale University Press, New Haven, CT.

Arrow, K. (1977). "Extended Sympathy and the Possibility of Social Choice". American Economic Review 67 (1), 219–225.

Arrow, K., Cline, W.R., Mäler, K., Munasinghe, M., Squitieri, R., Stiglitz, J. (1996a). "Intertemporal Equity, Discounting and Economic Efficiency". In: Bruce, J.P., Lee, H., Haites, E.F. (Eds.), Climate Change 1995: Economic and Social Dimensions of Climate Change. Cambridge University Press, Cambridge.

Arrow, K., Cropper, M., Eads, G., Hahn, R., Lave, L., Noll, R., Portney, P., Russell, M., Schmalensee, R., Smith, K., Stavins, R. (1996b). "Is There a Role for Benefit-Cost Analysis in Environmental, Health, and Safety Regulation?". Science 272, 221–222.

Arrow, K., Daily, G., Dasgupta, P.S., Ehrlich, P., Goulder, L., Heal, G.M., Levin, S., Mäler, K., Schneider, S., Starrett, D., Walker, B. (2004). "Are We Consuming Too Much?". Journal of Economic Perspectives 18 (3), 147–172.

Arrow, K., Solow, R., Portney, P.R., Leamer, E.E., Radner, R., Schuman, H. (1993). "Report of the NOAA Panel on Contingent Valuation". Federal Register 58 (10), 4601–4614.

Asheim, G., Buchholz, W., Tungodden, B. (2001). "Justifying Sustainability". Journal of Environmental Economics and Management 41 (3), 252–268.

Atkinson, S.E., Lewis, D.H. (1974). "A Cost-Effectiveness Analysis of Alternative Air Quality Control Strategies". Journal of Environmental Economics and Management 1 (3), 237–250.

Ayres, R.E. (2000). "Expanding the Use of Environmental Trading Programs Into New Areas of Environmental Regulation". Pace Environmental Law Review 18 (1), 87–118.

Bailey, E.M. (1996). "Allowance Trading Activity and State Regulatory Rulings: Evidence from the U.S. Acid Rain Program". MIT-CEEPR 96-002 working paper, Center for Energy and Environmental Policy Research, Massachusetts Institute of Technology, Cambridge, MA.

Barthold, T.A. (1994). "Issues in the Design of Environmental Excise Taxes". Journal of Economic Perspectives 8 (1), 133–151.

Bartik, T.J. (1988a). "Measuring the benefits of amenity improvements in hedonic price methods". Land Economics 64, 172–183.

Bartik, T.J. (1988b). "The Effects of Environmental Regulation on Business Location in the United States". Growth and Change 22 (Summer 1988).

Bartik, T. (1989). “Small Business Start-Ups in the United States: Estimates of the Effects of Characteristics of States”. South Economics Journal 55, 1004.

Bartik, T., Smith, V.K. (1987). “Urban Amenities and Public Policy”. In: Mills, E. (Ed.), Handbook of Urban Economics. North Holland, Amsterdam, pp. 1207–1254.

Baumol, W.J., Oates, W.E. (1988). The Theory of Environmental Policy. Cambridge University Press, Cambridge, U.K.

Becker, G.S. (1965). “A Theory of the Allocation of Time”. Economic Journal 75, 493–517.

Bergson, A. (1938). “A Reformulation of Certain Aspects of Welfare Economics”. Quarterly Journal of Economics 52 (2), 310–334.

Berman, E., Bui, L. (2001a). “Environmental Regulation and Labor Demand: Evidence from the South Coast Air Basin”. Journal of Public Economics 79 (2), 265–295.

Berman, E., Bui, L. (2001b). “Environmental Regulation and Productivity: Evidence from Oil Refineries”. Review of Economics and Statistics 83 (3), 498–510.

Bingham, T.H., Anderson, D.W., Cooley, P.C. (1987). “Distribution of the Generation of Air Pollution”. Journal of Environmental Economics and Management 14, 30–40.

Bishop, R.C. (1982). “Option value: An exposition and extension”. Land Economics 58 (1), 1–15.

Boardman, A., Greenberg, A., Vining, A., Weimar, D. (2001). Cost-Benefit Analysis: Concepts and Practice, 2nd edn. Prentice-Hall, Upper Saddle River, NJ.

Bockstael, N.E. (1996). “Travel Cost Methods”. In: Bromley, D. (Ed.), The Handbook of Environmental Economics. Blackwell Publishers, Oxford.

Bockstael, N.E., Hanemann, W.M., Strand, I.E. (1986). “Measuring the Benefits of Water Quality Improvements Using Recreation Demand Models”, Report prepared for the United States Environmental Protection Agency, Office of Policy Analysis, under Assistance Agreement #CR 811043.

Bockstael, N.E., McConnell, K.E. (1983). “Welfare Measurement in the Household Production Framework”. American Economic Review 73 (4), 806–814.

Bockstael, N.E., McConnell, K.E. (1993). “Public Goods as Characteristics of Non-Market Commodities”. Economic Journal 103, 1244–1257.

Bohi, D. (1994). “Utilities and State Regulators are Failing to Take Advantage of Emissions Allowance Trading”. The Electricity Journal 7 (2), 20–27.

Bohi, D., Burtraw, D. (1992). “Utility investment behavior and the emission trading market”. Resources and Energy 14, 129–153.

Bohi, D., Burtraw, D. (1997). “SO2 Allowance Trading: How do Expectations and Experience Measure Up”. The Electricity Journal 10 (7), 67–75.

Bohm, P. (1981). Deposit-Refund Systems: Theory and Applications to Environmental, Conservation, and Consumer Policy. Resources for the Future/Johns Hopkins University Press, Baltimore, MD.

Bohm, P., Russell, C.F. (1985). “Comparative Analysis of Alternative Policy Instruments”. In: Kneese, A.V., Sweeney, A.V. (Eds.), Handbook of Natural Resource and Energy Economics. North-Holland, Amsterdam.

Bovenberg, A.L., de Mooji, R. (1994). “Environmental Levies and Distortionary Taxation”. American Economic Review 84, 1085–1089.

Bovenberg, A.L., Goulder, L.H. (1996). “Optimal Environmental Taxation in the Presence of Other Taxes: General-Equilibrium Analyses”. American Economic Review 86, 985–1000.

Boyd, J., Ingberman, D.E. (1997). “The Search for Deep Pockets: Is ‘Extended Liability’ Expensive Liability?”. Journal of Law, Economics and Organization 13, 232–258.

Boyle, K.J. (2003). “Contingent Valuation in Practice”. In: Champ, P.A., Boyle, K.J., Brown, T.C. (Eds.), A Primer on Nonmarket Valuation. Kluwer Academic Publishers, Dordrecht.

Breyer, S. (1993). Breaking the Vicious Circle: Toward Effective Risk Regulation. Harvard University Press, Cambridge, MA.

Brown, J., Rosen, H. (1982). “On the Estimation of Structural Hedonic Price Models”. Econometrica 50, 765–769.

Brown, T.C. (2003). “Introduction to Stated Preference Methods”. In: Champ, P.A., Boyle, K.J., Brown, T.C. (Eds.), A Primer on Nonmarket Valuation. Kluwer Academic Publishers, Dordrecht.

Buchanan, J.M., Tullock, G. (1975). "Polluters' Profits and Political Response: Direct Control Versus Taxes". American Economic Review 65 (1), 139–147.

Burtraw, D. (1996). "The SO2 Emissions Trading Program: Cost Savings Without Allowance Trades". Contemporary Economic Policy 14, 79–94.

Burtraw, D., Krupnick, A.J., Mansur, E., Austin, D., Farrell, D. (1998). "The Costs and Benefits of Reducing Air Pollution Related to Acid Rain". Contemporary Economic Policy 16, 379–400.

Burtraw, D., Mansur, E. (1999). "The Environmental Effects of SO2 Trading and Banking". Environmental Science and Technology 33 (20), 3489–3494.

Cameron, T.A., Poe, G.L., Ethier, R.G., Schulze, W.D. (2002). "Alternative Non-market Value-Elicitation Methods: Are the Underlying Preferences the Same?". Journal of Environmental Economics and Management 44 (3), 391–421.

Carlson, C., Burtraw, D., Cropper, M., Palmer, K. (2000). "Sulfur Dioxide Control by Electric Utilities: What are the Gains from Trade?". Journal of Political Economy 108, 1292–1326.

Carraro, C., Soubeyran, A. (1996). "Environmental Policy and the Choice of Production Technology". In: Carraro, C., Katsoulacos, Y., Xepapadeas, A. (Eds.), Environmental Policy and Market Structure. Kluwer Academic Publishers, Dordrecht, pp. 151–180.

Carson, R.T., Flores, N.E., Hanemann, W.M. (1998). "Sequencing and Valuing Public Goods". Journal of Environmental Economics and Management 36 (3), 314–323.

Carson, R.T., Mitchell, R.C., Hanemann, M., Kopp, R.J., Presser, S., Ruud, P.A. (2003). "Contingent Valuation and Lost Passive Use: Damages from the Exxon Valdez Oil Spill". Environmental and Resource Economics 25 (3), 257–286.

Cason, T.N. (1995). "An Experimental Investigation of the Seller Incentives in EPA's Emission Trading Auction". American Economic Review 85, 905–922.

Champ, P.A., Boyle, K.J., Brown, T.C. (Eds.) (2003). A Primer on Nonmarket Valuation. Kluwer Academic Press, Boston.

Clawson, M., Knetsch, J.L. (1966). Economics of Outdoor Recreation. Johns Hopkins Press, Baltimore.

Coase, R. (1960). "The Problem of Social Cost". The Journal of Law and Economics 3 (1), 1–44.

Coates, D., Heid, V., Munger, M. (1994). "Not Equitable, Not Efficient: U.S. Policy on Low-Level Radioactive Waste Disposal". Journal of Policy Analysis and Management 13 (3), 526–538.

Coglianese, C., Marchant, G. (2004). "Shifting Sands: The Limits of Science in Setting Risk Standards". University of Pennsylvania Law Review 152 (4), 1255–1360.

Conrad, K. (2002). "Computable General Equilibrium Models in Environmental and Resource Economics". In: Tietenberg, T.H., Folmer, H. (Eds.), The International Yearbook of Environmental and Resource Economics 2002/2003: A Survey of Current Issues. Edward Elgar, Northampton, MA.

Courant, P.N., Porter, R. (1981). "Averting expenditure and the cost of pollution". Journal of Environmental Economics and Management 8, 321–329.

Court, A.T. (1939). "Hedonic Price Indexes with Automotive Examples". In: The Dynamics of Automobile Demand. The General Motors Corporation, New York, pp. 99–117.

Crandall, R.W. (1983). Controlling Industrial Pollution: The Economics and Politics of Clean Air. Brookings Institution, Washington, D.C.

Crandall, R.W., Gruenspecht, T.E., Keeler, T.E., Lave, L.B. (1986). Regulating the Automobile. The Brookings Institution, Washington, D.C.

Crocker, T.D. (1966). "The Structuring of Atmospheric Pollution Control Systems". In: Wolozin, H. (Ed.), The Economics of Air Pollution. W.W. Norton and Company, Inc., New York.

Cropper, M.L. (2000). "Has Economic Research Answered the Needs of Environmental Policy?". Journal of Environmental Economics and Management 39, 328–350.

Cropper, M.L., Aydede, S.K., Portney, P.R. (1992). "Public Preferences for Life Saving", Discussion Paper CRM 9201. Resources for the Future, Washington, D.C.

Cropper, M.L., Aydede, S.K., Portney, P.R. (1994). "Preferences for Life Saving Programs: How the Public Discounts Time and Age". Journal of Risk and Uncertainty 8, 243–265.

Cropper, M.L., Evans, W.N., Berardi, S.J., Ducla-Soares, M.M., Portney, P.R. (1992). "The Determinants of Pesticide Regulation: A Statistical Analysis of EPA Decision Making". Journal of Political Economy 100, 175–197.

Cropper, M.L., Freeman, A.M. (1991). "Environmental Health Effects". In: Braden, J.B., Kolstad, C.D. (Eds.), Measuring the Demand for Environmental Quality. Elsevier Science Publications, Amsterdam.

Cropper, M.L., Laibson, D. (1999). "The Implications of Hyperbolic Discounting for Project Evaluation". In: Portney, P.R., Weyant, J.P. (Eds.), Discounting and Intergenerational Equity. Resources for the Future, Washington, D.C.

Cropper, M.L., Oates, W.E. (1992). "Environmental Economics: A Survey". Journal of Economic Literature 30, 675–740.

Cropper, M.L., Sussman, F.G. (1990). "Valuing Future Risks to Life". Journal of Environmental Economics and Management 19, 160–174.

Cummings, R.G., Brookshire, D.S., Schulze, W.D. (1986). Valuing Environmental Goods: An Assessment of the Contingent Valuation Method. Rowman and Allanheld, Totowa, NJ.

Dales, J. (1968). Pollution, Property and Prices. University Press, Toronto.

Davis, R. (1963). "The Value of Outdoor Recreation: An Economic Study of the Maine Wood", Doctoral dissertation, Economics Department, Harvard University, Cambridge, MA.

DeLong, J. (1981). "Replies to 'Cost-Benefit Analysis: An Ethical Critique' ", AEI Journal on Government and Society Regulation (March/April), 39–40.

Denicolò, V. (1999). "Pollution-reducing Innovations Under Taxes or Permits". Oxford Economic Papers 51, 184–199.

Desvousges, W.H., Johnson, F.R., Banzhaf, H.S. (1998). Environmental Policy Analysis with Limited Information: Principles and Applications of the Transfer Method. Edward Elgar, Northampton, MA.

Desvousges, W.H., Johnson, F.R., Hudson, S.P., Gable, S.K., Ruby, M.C. (1996). Using Conjoint Analysis and Health-State Classifications to Estimate the Value of Health Effects of Air Pollution. Triangle Economic Research, Raleigh, NC.

Dewees, D.R., Sims, W.A. (1976). "The Symmetry of Effluent Charges and Subsidies for Pollution Control". Canadian Journal of Economics 9 (2), 323–331.

Diamond, P.A., Hausman, J.L. (1994). "Contingent Valuation: Is Some Number Better than No Number?". Journal of Economic Perspectives 8 (4), 45–64.

Doucet, J., Strauss, T. (1994). "On the Bundling of Coal Sulfur Dioxide Emissions Allowances". Energy Policy 22 (9), 764–770.

Downing, P.B., White, L.J. (1986). "Innovation in Pollution Control". Journal of Environmental Economics and Management 13, 18–27.

Eckstein, O. (1958). Water Resource Development: The Economics of Project Evaluation. Harvard University Press, Cambridge, MA.

Eckstein, O. (1961). "A Survey of the Theory of Public Expenditure Criteria". In: Buchanan, J.M. (Ed.), Public Finances: Needs, Sources and Utilization. Princeton University Press, Princeton, NJ.

Efaw, F., Lanen, W.N. (1979). "Impact of User Charges on Management of Household Solid Waste", Report prepared for the U.S. Environmental Protection Agency under Contract No. 68-3-2634, Princeton, N.J.

Ehrlich, I., Chuma, H. (1990). "A Model of the Demand for Longevity and the Value of Life Extension". Journal of Political Economy 98 (4), 761–782.

Ekeland, I., Heckman, J.J., Nesheim, L. (2002). "Identifying Hedonic Models". American Economic Review Proceedings 92, 304–309.

Ekeland, I., Heckman, J.J., Nesheim, L. (2004). "Identification and Estimation of Hedonic Models". Journal of Political Economy 112 (1), S60–S109.

El-Ashry, M.T., Gibbons, D.C. (1986). Troubled Waters: New Policies for Managing Water in the American West. World Resources Institute, Washington, D.C.

Ellerman, D. (2003). "The U.S. SO2 Cap-and-Trade Program". In: Proceedings of the OECD Workshop on Ex Post Evaluation of Tradeable Permits: Methodological and Policy Issues. OECD, Paris.

Ellerman, D., Joskow, P.L., Schmalensee, R., Montero, J.P., Bailey, E.M. (2000). Markets for Clean Air: The U.S. Acid Rain Program. Cambridge University Press, New York.

Ellerman, D., Montero, J. (1998). “The Declining Trend in Sulfur Dioxide Emissions: Implications for Allowance Prices”. Journal of Environmental Economics and Management 36, 26–45.

Ellerman, D., Schmalensee, R., Joskow, P., Montero, J., Bailey, E. (1997). Emissions Trading Under the U.S. Acid Rain Program: Evaluation of Compliance Costs and Allowance Market Performance. MIT Center for Energy and Environmental Policy Research, Cambridge, MA.

Elliott, E.D., Ackerman, B.A., Millian, J.C. (1985). “Toward a Theory of Statutory Evolution: The Federalization of Environmental Law”. Journal of Law, Economics and Organization 1, 313.

Engel, K.H. (1997). “State Environmental Standard-Setting: Is There a ‘Race’ and Is It ‘to the Bottom’?”. Hastings Law Journal 48, 271–378.

Engel, K.H., Rose-Ackerman, S. (2001). “Environmental Federalism in the United States: The Risks of Devolution”. In: Esty, D.C., Geradin, D. (Eds.), Regulatory Competition and Economic Integration: Comparative Perspectives. Oxford University Press, Oxford.

Esty, D.C. (1996). “Revitalizing Environmental Federalism”. Michigan Law Review 95, 570–653.

Esty, D.C., Geradin, D. (2001). “Regulatory Co-Opetition”. In: Esty, D.C., Geradin, D. (Eds.), Regulatory Competition and Economic Integration: Comparative Perspectives. Oxford University Press, Oxford.

Farrow, S. (1998). “Environmental Equity and Sustainability: Rejecting the Kaldor-Hicks Criteria”. Ecological Economics 27 (2), 183–188.

Ferrall, B.L. (1991). “The Clean Air Act Amendments of 1990 and the use of Market Forces to Control Sulfur Dioxide Emissions”. Harvard Journal on Legislation 28, 235–252.

Fischer, C., Parry, I.W.H., Pizer, W.A. (2003). “Instrument Choice for Environmental Protection When Technological Innovation is Endogenous”. Journal of Environmental Economics and Management 45, 523–545.

Fisher, A., Chestnut, I., Violette, D. (1989). “The Value of Reducing Risks of Death: A Note on New Evidence”. Journal of Policy Analysis and Management 8 (1), 88–100.

Fisher, A., McClelland, G.H., Schulze, W.D. (1988). “Measures of Willingness to Pay versus Willingness to Accept: Evidence, Explanations, and Potential Reconciliation”. In: Peterson, G.L., Driver, B.L., Gregory, R. (Eds.), Amenity Resource Valuation: Integrating Economics with Other Disciplines. Venture, State College, PA, pp. 127–134.

Fisher, B., Barrett, S., Bohm, P., Kuroda, M., Mubazi, J., Shah, A., Stavins, R. (1996). “Policy Instruments to Combat Climate Change”. In: Bruce, J.P., Lee, H., Haites, E.F. (Eds.), Climate Change 1995: Economic and Social Dimensions of Climate Change. Cambridge University Press, New York, pp. 397–439.

Foster, V., Hahn, R.W. (1995). “Designing More Efficient Markets: Lessons from Los Angeles Smog Control”. Journal of Law and Economics 38, 19–48.

Frankel, J.A. (1999). “Greenhouse Gas Emissions”. The Brookings Institution, Washington, D.C.

Frederick, K.D. (Ed.) (1986). Scarce Water and Institutional Change. Resources for the Future, Washington, D.C.

Freeman, A.M. III (1997). “On Valuing the Services and Functions of Ecosystems”. In: Simpson, D.R., Christensen Jr., N.L. (Eds.), Ecosystem Function and Human Activities: Reconciling Economics and Ecology. Chapman and Hall, New York.

Freeman, A.M. III (2003). The Measurement of Environmental and Resource Values: Theory and Methods, 2nd edn. Resources for the Future, Washington, D.C.

Freeman, A.M. III, Haveman, R.H. (1972). “Clean Rhetoric and Dirty Water”. Public Interest 28, 51–65.

Friedman, J., Gerlowski, D.A., Silberman, J. (1992). “What Attracts Foreign Multinational Corporations? Evidence from Branch Plant Locations in the United States”. Journal of Regulatory Science 32, 403.

Fullerton, D., Kinnaman, T.C. (1996). “Household Responses to Pricing Garbage by the Bag”. American Economic Review 86, 971–984.

Fullerton, D., Stavins, R. (1998). “How Economists See the Environment”. Nature 395 (10), 433–434.

Gianessi, L.P., Peskin, H.M., Wolff, E. (1979). “The Distributional Effects of Uniform Air Pollution Policy in the United States”. Quarterly Journal of Economics 93, 281–301.

Goklany, I.M. (1998a). “Did Federalization Halt a Race to the Bottom for Air Quality?”. Environmental Manager (June), 12–17.

Goklany, I.M. (1998b). "Do We Need the Federal Government to Protect Air Quality?". Center for the Study of American Business Policy, Study No. 150.

Goldberg, P.K. (1998). "The Effects of the Corporate Average Fuel Efficiency Standards in the US". Journal of Industrial Economics 46 (1), 1–33.

Goldstein, J.B. (1991). "The Prospects for Using Market Incentives to Conserve Biological Diversity". Environmental Law 21 (3), 985–1014.

Goulder, L.H. (1995a). "Effects of Carbon Taxes in an Economy with Prior Tax Distortions: An Intertemporal General Equilibrium Analysis". Journal of Environmental Economics and Management 29, 271–297.

Goulder, L.H. (1995b). "Environmental Taxation and the Double Dividend: A Reader's Guide". International Tax and Public Finance 2, 157–183.

Goulder, L.H., Perry, I.W.H., Burtraw, D. (1997). "Revenue-Raising versus Other Approaches to Environmental Protection: The Critical Significance of Pre-Existing Tax Distortions". RAND Journal of Economics 28 (4), 708–731.

Goulder, L.H., Perry, I.W.H., Williams, R.C.I., Burtraw, D. (1999). "The Cost-Effectiveness of Alternative Instruments for Environmental Protection in a Second Best Setting". Journal of Public Economics 72 (3), 329–360.

Goulder, L.H., Stavins, R.N. (2002). "An Eye on the Future: How Economists' Controversial Practice of Discounting Really Affects the Evaluation of Environmental Policies". Nature 419 (10), 673–674.

Graham, D.A. (1981). "Cost-Benefit Analysis under Uncertainty". American Economic Review 71 (4), 715–725.

Graham, J.D. (2001). Memorandum for the President's Management Council, "Presidential Review of Agency Rulemaking by OIRA", 20 September.

Graham, J.D., Loevzel, K. (1997). "Regulatory Reform: Moving Forward in the States". Risk in Perspective 5 (2), 1–2. Harvard Center for Risk Analysis.

Graham, J.D., Wiener, J.B. (1995). Risk versus Risk: Tradeoffs in Protecting Health and the Environment. Harvard University Press, Cambridge, MA.

Griliches, Z. (1961). "Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change". In: The Price Statistics of the Federal Government. Columbia University Press for the National Bureau of Economic Research, New York, pp. 137–196.

Grossman, M. (1972). "On the concept of health capital and the demand for health". Journal of Political Economy 80, 223–255.

Gruenspecht, H.K. (1981). "Differentiated Social Regulation in Theory and Practice", Ph.D. dissertation, Yale University.

Gruenspecht, H.K. (1982). "Differentiated Regulation: The Case of Auto Emissions Standards". The American Economic Review Papers and Proceedings 72 (2), 328–331.

Gupta, S., Van Houtven, G., Cropper, M. (1996). "Paying for permanence: an economic analysis of EPA's cleanup decisions at Superfund sites". Rand Journal of Economics 27 (3), 563–582.

Haab, T.C., McConnell, K.E. (2002). Valuing Environmental and Natural Resources: The Econometrics of Non-market Valuation. Edward Elgar, Northampton, MA.

Hahn, R.W. (1984). "Market Power and Transferable Property Rights". Quarterly Journal of Economics 99 (4), 753–765.

Hahn, R.W. (1987). "Jobs and Environmental Quality: Some Implications for Instrument Choice". Policy Sciences 20, 289–306.

Hahn, R.W. (1989). "Economic Prescriptions for Environmental Problems: How the Patient Followed the Doctor's Orders". Journal of Economic Perspectives 3, 95–114.

Hahn, R.W. (2000). "The Impact of Economics on Environmental Policy". Journal of Environmental Economics and Management 39 (3), 375–399.

Hahn, R.W., Axtell, R.L. (1995). "Reevaluating the Relationship between Transferable Property Rights and Command-and-Control Regulation". Journal of Regulatory Economics 8 (2), 125–148.

Hahn, R.W., Dudley, P. (2004). "How Well Does the Government Do Cost-Benefit Analysis?", Working Paper 04-01, AEI-Brookings Joint Center for Regulatory Studies, Washington, D.C.

Hahn, R.W., Hester, G.L. (1989). “Marketable permits: lessons for theory and practice”. Ecological Law Quarterly 16 (2), 361–406.

Hahn, R.W., Lutter, R.W., Viscusi, K. (2000). Do Federal Regulations Reduce Mortality? AEI-Brookings Joint Center for Regulatory Studies, Washington, D.C.

Hahn, R.W., McGartland, A.M. (1989). “Political Economy of Instrument Choice: An Examination of the U.S. Role in Implementing the Montreal Protocol”. Northwestern University Law Review 83, 592–611.

Hahn, R.W., Noll, R.G. (1982). “Designing a Market for Tradeable Permits”. In: Magat, W. (Ed.), Reform of Environmental Regulation. Ballinger Publishing, Co., Cambridge, MA.

Hahn, R.W., Noll, R.G. (1990). “Environmental Markets in the Year 2000”. Journal of Risk and Uncertainty 3 (4), 351–367.

Hahn, R.W., Olmstead, S.M., Stavins, R. (2003). “Environmental Regulation During the 1990’s: A Retrospective Analysis”. Harvard Environmental Law Review 27 (2), 377–415.

Hahn, R.W., Stavins, R.N. (1991). “Incentive-based Environmental Regulation: a New Era from an Old Idea?”. Ecology Law Quarterly 18, 1–42.

Hahn, R.W., Stavins, R.N. (1992). “Economic Incentives for Environmental Protection: Integrating Theory and Practice”. American Economic Review 82, 464–468.

Hahn, R.W., Stavins, R.N. (1995). “Trading in Greenhouse Permits: A Critical Examination of Design and Implementation Issues”. In: Lee, H. (Ed.), Shaping National Responses to Climate Change: A Post-Rio Policy Guide. Island Press, Cambridge, MA, pp. 177–217.

Hamilton, J. (1993). “Politics and Social Costs: Estimating the Impact of Collective Action on Hazardous Waste Facilities”. Rand Journal of Economics 24, 101–125.

Hamilton, J. (1995). “Pollution as News: Media and Stock Market Reactions to the Toxics Release Inventory Data”. Journal of Environmental Economics and Management 28, 98–113.

Hamilton, J.T., Viscusi, K. (1999). Calculating Risks? The Spatial and Political Dimensions of Hazardous Waste Policy. MIT Press, Cambridge, MA.

Hammitt, J.K. (2000). “Are the Costs of Proposed Environmental Regulations Overestimated? Evidence from the CFC Phaseout”. Environmental and Resource Economics 16 (3), 281–301.

Hanneman, M. (1991). “Willingness to pay and willingness to accept: how much can they differ?”. American Economic Review 81 (3), 635–647.

Hanneman, M. (1994). “Valuing the Environment Through Contingent Value”. Journal of Economic Perspectives 8 (4), 19–43.

Harberger, A.C. (1978). “On the Use of Distributional Weights in Social Cost-Benefit Analysis”. Journal of Political Economy 86, 87–120.

Harrington, W., Morgenstern, R.D. (2003). “International Experience with Competing Approaches to Environmental Policy: Results from Six Paired Cases”, Working paper, July 2003, Resources for the Future, Washington, D.C.

Harrington, W., Morgenstern, R.D., Nelson, P. (2000). “On the Accuracy of Regulatory Cost Estimates”. Journal of Policy Analysis and Management 19 (2), 297–322.

Harrington, W., Portney, P.R. (1987). “Valuing the benefits of health and safety regulations”. Journal of Urban Economics 22, 101–112.

Harrison, D. Jr. (1999). “Turning Theory into Practice for Emissions Trading in the Los Angeles Air Basin”. In: Sorrell, S., Skea, J. (Eds.), Pollution for Sale: Emissions Trading and Joint Implementation. Edward Elgar, London.

Harrison, D. Jr. (2003). “Ex Post Evaluation of the RECLAIM Emissions Trading Program for the Los Angeles Air Basin”. In: Proceedings of the OECD Workshop on Ex Post Evaluation of Tradeable Permits: Methodological and Policy Issues. OECD, Paris.

Hartman, R.S., Bozdogan, K., Nadkarni, R.M. (1979). “The Economic Impact of Environmental Regulations on the U.S. Copper Industry”. The Bell Journal of Economics 10, 589–618.

Hartman, R.S., Wheeler, D., Singh, M. (1994). The Cost of Air Pollution Abatement, World Bank Policy Research Working Paper #1398, World Bank, Washington, D.C.

Hartridge, O. (2003). “The UK Emissions Trading Scheme: A Progress Report”. In: Proceedings of the OECD Workshop on Ex Post Evaluation of Tradeable Permits: Methodological and Policy Issues. OECD, Paris.

Hausman, J.A. (Ed.) (1993). Contingent Valuation: A Critical Assessment. North Holland, Amsterdam.

Haveman, R. (1965). Water Resource Investment and the Public Interest: An Analysis of Federal Expenditure in Ten Southern States. Vanderbilt University Press, Nashville, TN.

Hazilla, M., Kopp, R.J. (1990). "Social cost of environmental quality regulations: A general equilibrium analysis". Journal of Political Economy 98 (4), 853–873.

Heinzerling, L. (1998). "Regulatory Costs of Mythic Proportions". Yale Law Journal 107, 1981–2070.

Heinzerling, L. (2001). "The Clean Air Act and the Constitution". St. Louis University Public Law Review 20, 121–151.

Herriges, J.A., Kling, C.L. (Eds.) (1999). Valuing Recreation and the Environment. Edward Elgar, Northampton, MA.

Herriges, J.A., Kling, C.L., Phaneuf, D.J. (2004). "What's the use? Welfare estimates from revealed preference models when weak complementarity does not hold". Journal of Environmental Economics and Management 47 (1), 55–70.

Hicks, J.R. (1939). "The Foundations of Welfare Economics". The Economic Journal 49, 696–712.

Hird, J.A. (1993). "Environmental Policy and Equity: The Case of Superfund". Journal of Policy Analysis and Management 12 (2), 323–343.

Hockenstein, J.B., Stavins, R.N., Whitehead, B.W. (1997). "Crafting the Next Generation of Market-Based Environmental Tools". Environment 39 (4), 12–20, 30–33.

Horowitz, J.K., Carson, R.T. (1990). "Discounting Statistical Lives". Journal of Risk and Uncertainty 3, 402–413.

Horowitz, J., McConnell, K.E. (2002). "A Review of WTA/WTP Studies". Journal of Environmental Economics and Management 44, 426–447.

Horowitz, J., McConnell, K.E. (2003). "Willingness to Accept, Willingness to Pay, and the Income Effect". Journal of Economic Behavior and Organization 51, 537–545.

Hotelling, H. (1947). "Letter to the National Park Service". In: An Economic Study of the Monetary Evaluation of Recreation in the National Parks. U.S. Department of the Interior, National Park Service and Recreational Planning Division, Washington, DC, 1949.

Howe, C.W. (1997). "Increasing Efficiency in Water Markets: Examples from the Western United States". In: Anderson, T.L., Hill, P.J. (Eds.), Water Marketing—The Next Generation. Rowman and Littlefield Publishers, Inc., Lanham, MD, pp. 79–99.

Hylland, A., Zeckhauser, R.J. (1979). "Distributional Objectives Should Affect Taxes But Not Program Choice or Design". Scandinavian Journal of Economics 81 (2), 264–284.

Jaffe, A.B., Newell, R.G., Stavins, R.N. (2002). "Environmental Policy and Technological Change". Environmental and Resource Economics 22, 41–69.

Jaffe, A.B., Newell, R.G., Stavins, R.N. (2003). "Technological Change and the Environment". In: Mäler, K.G., Vincent, J. (Eds.), The Handbook of Environmental Economics, vol. I. North Holland/Elsevier Science, Amsterdam.

Jaffe, A.B., Peterson, S.R., Portney, P.R., Stavins, R.N. (1995). "Environmental Regulation and the Competitiveness of U.S. Manufacturing: What Does the Evidence Tell Us?". Journal of Economic Literature 33, 132–165.

Jaffe, A.B., Stavins, R.N. (1995). "Dynamic Incentives of Environmental Regulation: The Effects of Alternative Policy Instruments on Technological Diffusion". Journal of Environmental Economics and Management 29, S43–S63.

Johansson, P.-O. (1995). Evaluating Health Risks: An Economic Approach. Cambridge University Press, Cambridge, U.K.

Johnson, S.L., Pekelney, D.M. (1996). "Economic Assessment of the Regional Clean Air Incentives Market: A New Emissions Trading Program for Los Angeles". Land Economics 72, 277–297.

Jones-Lee, M.W. (1974). "The Value of Changes in the Probability of Death or Injury". Journal of Political Economy 82, 835–849.

Jones-Lee, M.W., Hammerton, M., Philips, P.R. (1985). "The Value of Safety: Results of a National Sample Survey". Economic Journal 95, 49–72.

Jorgenson, D., Wilcoxen, P. (1994). "The Economic Effects of a Carbon Tax", Paper presented to the IPCC Workshop on Policy Instruments and their Implications, January 17–20, Tsukuba, Japan.

Jorgenson, D.W. (1997). Welfare: Measuring Social Welfare, vol. 2. MIT Press, Cambridge, MA.

Joskow, P.L., Schmalensee, R. (1998). "The Political Economy of Market-based Environmental Policy: The U.S. Acid Rain Program". Journal of Law and Economics 41, 81–135.

Joskow, P.L., Schmalensee, R., Bailey, E.M. (1998). "Auction Design and the Market for Sulfer Dioxide Emissions". American Economic Review 88 (4), 669–685.

Jung, C., Krutilla, K., Boyd, R. (1996). "Incentives for Advanced Pollution Abatement Technology at the Industry Level: An Evaluation of Policy Alternatives". Journal of Environmental Economics and Management 30, 95–111.

Kaldor, N. (1939). "Welfare Propositions of Economics and Interpersonal Comparisons of Utility". Economic Journal 49, 549–552.

Kaplow, L. (1996). "The Optimal Supply of Public Goods and the Distortionary Cost of Taxation". National Tax Journal 49 (4), 513–533.

Kaplow, L. (2004). "On the (Ir)relevance of Distribution and Labor Supply Distortion to Government Policy". Journal of Economic Perspectives 18 (4), 159–175.

Kaplow, L., Shavell, S. (1994). "Why the Legal System is Less Efficient Than the Income Tax in Redistributing Income". Journal of Legal Studies 23 (2), 667–681.

Kaplow, L., Shavell, S. (2001). "Any Non-welfarist Method of Policy Assessment Violates the Pareto Principle". Journal of Political Economy 109 (2), 281–286.

Kaplow, L., Shavell, S. (2002a). Fairness versus Welfare. Harvard University Press, Cambridge, MA.

Kaplow, L., Shavell, S. (2002b). "On the Superiority of Corrective Taxes to Quantity Regulation". American Law and Economics Review 4 (1), 1–17.

Katsoulacos, Y., Xepapadeas, A. (1996). "Environmental Innovation, Spillovers and Optimal Policy Rules". In: Carraro, C., Katsoulacos, Y., Xepapadeas, A. (Eds.), Environmental Policy and Market Structure. Kluwer Academic Publishers, Dordrecht, pp. 143–150.

Keeler, A.G. (1991). "Noncompliant firms in transferable discharge permit markets: Some extensions". Journal of Environmental Economics and Management 21, 180–189.

Keeney, R.L. (1990). "Mortality Risks Induced by Economic Expenditures". Risk Analysis 10, 147–159.

Keeney, R.L. (1997). "Estimating Fatalities Induced by the Economic Costs of Regulations". Journal of Risk and Uncertainty 14 (1), 5–23.

Kelman, S. (1981a). "Cost-Benefit Analysis: An Ethical Critique". AEI Journal on Government and Society Regulation (January/February), 33–40.

Kelman, S. (1981b). What Price Incentives?: Economists and the Environment. Auburn House, Boston.

Kemp, R., Soete, L. (1990). "Inside the 'Green Box': On the Economics of Technological Change and the Environment". In: Freeman, C., Soete, L. (Eds.), New Explorations in the Economics of Technological Change. Pinter, London.

Keohane, N.O. (1999). "Policy Instruments and the Diffusion of Pollution Abatement Technology". Working paper, Harvard University.

Keohane, N.O. (2001). "Essays in the Economics of Environmental Policy". Unpublished Ph.D. thesis, Harvard University.

Keohane, N.O. (2003). "What Did the Market Buy? Cost Savings Under the U.S. Tradeable Permits Program for Sulfur Dioxide". Yale Center for Environmental Law and Policy Working Paper 01-11-2003, Yale School of Management.

Keohane, N.O., Revesz, R.L., Stavins, R.N. (1998). "The Choice of Regulatory Instruments in Environmental Policy". Harvard Environmental Law Review 22 (2), 313–367.

Kerr, S., Maré, D. (1997). "Efficient Regulation Through Tradeable Permit Markets: The United States Lead Phasedown". Working Paper 96-06, Department of Agricultural and Resource Economics, University of Maryland, College Park.

Kerr, S., Newell, R. (2003). "Policy-Induced Technology Adoption: Evidence from the U.S. Lead Phasedown". Journal of Industrial Economics 51 (3), 317–343.

Kling, C., Rubin, J.D. (1997). "Bankable Permits for Control of Environmental Pollution". Journal of Public Economics 64 (1), 99–113.

Kneese, A.V., Schulze, C.L. (1975). Pollution, Prices, and Public Policy. Brookings Institution, Washington, D.C.

Kohn, R.E. (1985). "A General Equilibrium Analysis of the Optimal Number of Firms in a Polluting Industry". Canadian Journal of Economics 18 (2), 347–354.

Kolstad, C. (1986). "Empirical Properties of Economic Incentives and Command-and-Control Regulations for Air Pollution Control". Land Economics 62, 250–268.

Kolstad, C., Ulen, T., Johnson, G. (1990). "*Ex Post* Liability for Harm vs. *Ex Ante* Safety Regulation: Substitutes or Complements?". American Economic Review 80 (4), 888–901.

Konar, S., Cohen, M.A. (1997). "Information as Regulation: The Effect of Community Right to Know Laws on Toxic Emissions". Journal of Environmental Economics and Management 32, 109–124.

Koplow, D.N. (1993). Federal Energy Subsidies: Energy, Environmental and Fiscal Impacts. Alliance to Save Energy, Lexington, MA.

Kopp, R.J. (1993). "Environmental Economics: Not Dead but Thriving". Resources (Spring), 7–12.

Kopp, R.J., Krupnick, A.J., Toman, M. (1997). "Cost-Benefit Analysis and Regulatory Reform: An Assessment of the Science and the Art". Discussion Paper 97-19, Resources for the Future, Washington, D.C.

Kornhauser, L.A., Revesz, R.L. (1989). "Sharing Damages Among Multiple Tortfeasors". Yale Law Journal 98, 831–884.

Kornhauser, L.A., Revesz, R.L. (1990). "Apportioning Damages Among Potentially Insolvent Actors". Journal of Legal Studies 19, 617–651.

Kornhauser, L.A., Revesz, R.L. (1994a). "Multidefendant Settlements: The Impact of Joint and Several Liability". Journal of Legal Studies 23, 41–76.

Kornhauser, L.A., Revesz, R.L. (1994b). "Multidefendant Settlements Under Joint and Several Liability: The Problem of Insolvency". Journal of Legal Studies 23, 517–542.

Kornhauser, L.A., Revesz, R.L. (1995). "Evaluating the Effects of Alternative Superfund Liability Rules". In: Revesz, R., Stewart, R. (Eds.), Analyzing Superfund: Economics, Science, and Law. Resources for the Future, Washington, D.C.

Kornhauser, L.A., Revesz, R.L. (2000). "Joint Tortfeasors". In: Bouckaert, B., Geest, G.D. (Eds.), Encyclopedia of Law and Economics: Civil Law and Economics, vol. II. Edward Elgar Publishing, Inc., Cheltenham, UK.

Krier, J.E. (1995). "On the Topology of Uniform Environmental Standards in a Federal System—And Why It Matters". Modern Law Review 54, 1226.

Krupnick, A., Alberini, A., Cropper, M., Simon, N., O'Brien, B., Goeree, R., Heintzelman, M. (2002). "Age, Health and the Willingness to Pay for Mortality Risk Reductions: A Contingent Valuation Survey of Ontario Residents". Journal of Risk and Uncertainty 24 (2), 161–186.

Kuhn, H.W., Tucker, A.W. (1951). "Nonlinear programming". In: Neyman, J. (Ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, pp. 481–492.

Lave, L. (1981). The Strategy of Social Regulation: Decision Frameworks for Policy. Brookings Institution, Washington, D.C.

Lave, L., Gruenspecht, H. (1991). "Increasing the Efficiency and Effectiveness of Environmental Decisions: Benefit-Cost Analysis and Effluent Fees". Journal of Air and Waste Management 41, 680–690.

Leontief, W.W. (1966). Input-Output Economics. Oxford University Press, New York.

Leontief, W.W. (1970). "Environmental repercussions and the economic structure: an input-output approach". Review of Economics and Statistics 38, 386–407.

Levinson, A. (1996). "Environmental Regulations and Manufacturers' Location Choices: Evidence from the Census of Manufacturers". Journal of Public Economics 62, 5.

Lind, R.C. (1982). "A Primer on the Major Issues Relating to the Discount Rate for Evaluating National Energy Options". In: Lind, R.C., Arrow, K.J., Corey, G.R., Dasgupta, P., Sen, A.K., Stauffer, T., Stiglitz, J.E., Stockfisch, J.A., Wilson, R. (Eds.), Discounting for Time and Risk in Energy Policy. Resources for the Future, Washington, D.C.

Lind, R.C. (1995). “Intergenerational Equity, Discounting, and the Role of Cost-Benefit Analysis in Evaluating Global Climate Policy”. Energy Policy 23 (4–5), 379–389.

Lowry, R.C. (1993). “The Political Economy of Environmental Citizen Groups”. Unpublished Ph.D. thesis, Harvard University.

Lutter, R. (1999). “Is EPA’s Ozone Standard Feasible?”. In: Regulatory Analysis 99-6. AEI-Brookings Joint Center for Regulatory Studies, Washington, D.C. (December).

Lutter, R.W., Morrall, J.F. (1994). “Health-Health Analysis: A New Way to Evaluate Health and Safety Regulation”. Journal of Risk and Uncertainty 8 (1), 43–66.

MacDonnell, L.J. (1990). The Water Transfer Process As a Management Option for Meeting Changing Water Demands, vol. I. Submitted to the U.S. Geological Survey, Washington, D.C.

Magat, W.A. (1978). “Pollution control and technological advance: A dynamic model of the firm”. Journal of Environmental Economics and Management 5, 1–25.

Magat, W.A. (1979). “The Effects of Environmental Regulation on Innovation”. Law and Contemporary Problems 43, 3–25.

Mäler, K.G. (1985). “Welfare Economics and the Environment”. In: Kneese, A.V., Sweeney, J.L. (Eds.), Handbook of Natural Resource and Energy Economics. North Holland, Amsterdam.

Maleug, D.A. (1989). “Emission Credit Trading and the Incentive to Adopt New Pollution Abatement Technology”. Journal of Environmental Economics and Management 16, 52–57.

Maleug, D.A. (1990). “Welfare consequences of emission credit trading programs”. Journal of Environmental Economics and Management 18, 66–77.

Maloney, M., Brady, G.L. (1988). “Capital Turnover and Marketable Property Rights”. The Journal of Law and Economics 31 (1), 203–226.

Maloney, M.T., McCormick, R.E. (1982). “A Positive Theory of Environmental Quality Regulation”. Journal of Law and Economics 25 (1), 99–124.

Maloney, M.T., Yandle, B. (1984). “Estimation of the Cost of Air Pollution Control Regulation”. Journal of Environmental Economics and Management 11, 244–263.

Marin, A. (1991). “Firm Incentives to Promote Technological Change in Pollution Control: Comment”. Journal of Environmental Economics and Management 21, 297–300.

Markusen, J.R., Morey, E.R., Olewiler, N.D. (1993). “Environmental Policy when Market Structure and Plant Location are Endogenous”. Journal of Environmental Economics and Management 24, 69.

Markusen, J.R., Morey, E.R., Olewiler, N.D. (1995). “Competition in Regional Environmental Policies when Plant Locations are Endogenous”. Journal of Public Economics 56, 55.

McConnell, V.D., Schwab, R.M. (1991). “The Impact of Environmental Regulation on Industry Location Decisions: The Motor Vehicle Industry”. Land Economics 66, 67.

McCubbins, M.D., Noll, R.G., Weingast, B.R. (1989). “Structure and Process, Politics and Policy: Administrative Arrangements and the Political Control of Agencies”. Virginia Law Review 75 (2), 431–482.

McCubbins, M.D., Sullivan, T. (1984). “Constituency Influences on Legislative Policy Choice”. Quality and Quantity 18, 299–319.

McFarland, J.M. (1972). “Economics of Solid Waste Management”. In: Comprehensive Studies of Solid Waste Management, Final Report, No. 72-3. Sanitary Engineering Research Laboratory, University of California, Berkeley, pp. 41–106.

Mendelsohn, R. (1986). “Regulating Heterogeneous Emissions”. Journal of Environmental Economics and Management 13 (4), 301–312.

Menell, P.S. (1990). “Beyond the Throwaway Society: An Incentive Approach to Regulating Municipal Solid Waste”. Ecology Law Quarterly 17, 655–739.

Menell, P.S. (1991). “The Limitations of Legal Institutions For Addressing Environmental Risks”. Journal of Economic Perspectives 5, 93–113.

Menell, P.S. (2002). Environmental Law, International Library of Essays in Law and Legal Theory, Second Series. Ashgate Publishing, Aldershot, United Kingdom.

Menell, P.S. (2003). “The Municipal Solid Waste ‘Crisis’ in Retrospect: A Success Story for Market-Based Mechanisms”. Paper prepared for presentation at “Twenty Years of Market-Based Instruments for Environmental Protection: Has the Promise Been Realized?”, Donald Bren School of Environmental Science and Management, University of California, Santa Barbara, August 23–24, 2003, U.C. Berkeley.

Mestelman, S. (1982). "Production Externalities and Corrective Subsidies: A General Equilibrium Analysis". Journal of Environmental Economics and Management 9 (2), 186–193.

Miller, T.R. (1989). Narrowing the Plausible Range Around the Value of Life. The Urban Institute, Washington, D.C.

Milliman, S.R., Prince, R. (1989). "Firm Incentives to Promote Technological Change in Pollution Control". Journal of Environmental Economics and Management 17, 247–265.

Milliman, S.R., Prince, R. (1992). "Firm Incentives to Promote Technological Change in Pollution Control: Reply". Journal of Environmental Economics and Management 22, 292–296.

Miranda, M.L., Everett, J.W., Blume, D., Roy, B.A. (1994). "Market-Based Incentives and Residential Municipal Solid Waste". Journal of Policy Analysis and Management 13, 681–698.

Mishan, E. (1976). Cost-Benefit Analysis. Praeger Publishers, New York.

Misolek, W.S., Elder, H.W. (1989). "Exclusionary Manipulation of Markets for Pollution Rights". Journal of Environmental Economics and Management 16, 156–166.

Mitchell, R.C., Carson, R.T. (1989). Using Surveys to Value Public Goods: The Contingent Valuation Method. Resources for the Future, Washington, D.C.

Montero, J.P. (2002). "Market Structure and Environmental Innovation". Journal of Applied Economics 5, 293–325.

Montero, J.P. (2003). "Tradeable Permits with Imperfect Monitoring: Theory and Evidence". Working paper, MIT Center for Energy and Environmental Policy.

Montgomery, D.W. (1972). "Markets in Licenses and Efficient Pollution Control Programs". Journal of Economic Theory 5, 395.

Moore, M.J., Viscusi, W.K. (1988). "The Quantity-Adjusted Value of Life". Economic Inquiry 26 (3), 369–388.

Morall, J.F. (2003). "Saving Lives: A Review of the Record". Journal of Risk and Uncertainty 27 (3), 221–237.

Morgenstern, R.D. (2000). "Decision making at EPA: Economics, Incentives and Efficiency", Draft conference paper, "EPA at Thirty: Evaluating and Improving the Environmental Protection Agency", 7–8 December, Duke University.

Morgenstern, R.D. (1997). "The Legal and Institutional Setting for Economic Analysis at EPA". In: Morgenstern, R.D. (Ed.), Economic Analyses at EPA: Assessing Regulatory Impact. Resources for the Future, Washington, D.C, pp. 5–23.

Morris, J., Scarlett, L. (1996). "Buying Green: Consumers, Product Labels and the Environment". Policy Study No. 202, The Reason Foundation, Los Angeles, CA.

Mrozek, J., Taylor, L.O. (2002). "What Determines the Value of Life? A Meta Analysis". Journal of Policy Analysis and Management 21 (2), 253–270.

Nash, J.R., Revesz, R.L. (2001). "Markets and Geography: Designing Marketable Permit Schemes to Control Local and Regional Pollutants". Ecology Law Quarterly 28, 569–661.

National Research Council (2002). "Effectiveness and Impact of Corporate Average Fuel Economy (CAFE) Standards". Committee on the Effectiveness and Impact of Corporate Average Fuel Economy (CAFE) Standards, Board on Energy and Environmental Systems, Transportation Research Board. The National Academies Press, Washington, D.C.

Nelson, R., Tietenberg, T.H., Donihue, M. (1993). "Differential Environmental Regulation: Effects on Electric Utility Capital Turnover and Emissions". Review of Economics and Statistics 75 (2), 368–373.

Newell, R.G., Jaffe, A.B., Stavins, R.N. (1999). "The Induced Innovation Hypothesis and Energy-Saving Technological Change". Quarterly Journal of Economics 114 (3), 941–975.

Newell, R.G., Pizer, W.A. (2003a). "Discounting the Distant Future: How Much Do Uncertain Rates Increase Valuations?". Journal of Environmental Economics and Management 46 (1), 52–71.

Newell, R.G., Pizer, W.A. (2003b). "Regulating Stock Externalities Under Uncertainty". Journal of Environmental Economics and Management 45, 416–432.

Newell, R.G., Pizer, W.A. (2004). "Uncertain Discount Rates in Climate Policy Analysis". Energy Policy 32 (4), 519–529.

Newell, R.G., Stavins, R.N. (2003). "Cost Heterogeneity and the Potential Savings from Market-Based Policies". Journal of Regulatory Economics 23 (1), 43–59.

Nichols, A. (1997). "Lead in Gasoline". In: Morgenstern, R.D. (Ed.), Economic Analyses at EPA: Assessing Regulatory Impact. Resources for the Future, Washington, D.C.

Nwaneri, V.C. (1970). "Equity in Cost-Benefit Analysis: A Case Study of the Third London Airport". Journal of Transportation Economics and Policy 4 (3), 235–254.

Oates, W.E., Schwab, R.M. (1988). "Economic Competition Among Jurisdictions: Efficiency Enhancing or Distortion Inducing?". Journal of Public Economics 35, 333.

Organization for Economic Cooperation and Development (1989). Economic Instruments for Environmental Protection. OECD, Paris.

Organization for Economic Cooperation and Development (1991). Environmental Policy: How to Apply Economic Instruments. OECD, Paris.

Organization for Economic Cooperation and Development (1998). Applying Market-Based Instruments to Environmental Policies in China and OECD Countries. OECD, Paris.

Orr, L. (1976). "Incentive for Innovation as the Basis for Effluent Charge Strategy". American Economic Review 66, 441–447.

Palmer, K., Oates, W.E., Portney, P.R. (1995). "Tightening Environmental Standards: The Benefit-Cost or the No-Cost Paradigm?". Journal of Economic Perspectives 9, 119–132.

Palmquist, R.B. (1991). "Hedonic Methods". In: Braden, J., Kolstad, C. (Eds.), Measuring the Demand for Environmental Quality. Elsevier Science, Amsterdam.

Palmquist, R.B. (2005). "Weak Complementarity, Path Independence, and the Intuition of the Willig Condition". Journal of Environmental Economics and Management 49 (1), 103–115.

Palmquist, R.B. (2005). "Property Value Models". In: Mäler, K., Vincent, J. (Eds.), Handbook of Environmental Economics, vol. 2. North Holland/Elsevier Science, Amsterdam.

Pareto, V. (1896). Cours d'Economie Politique, vol. 2. Lausanne.

Parry, I. (1995). "Pollution, Taxes, and Revenue Recycling". Journal of Environmental Economics and Management 29, 64–77.

Parry, I.W.H. (1998). "Pollution Regulation and the Efficiency Gains from Technological Innovation". Journal of Regulatory Economics 14, 229–254.

Parsons, G.R. (2003). "The Travel Cost Model". In: Champ, P.A., Boyle, K.J., Brown, T.C. (Eds.), A Primer on Nonmarket Valuation. Kluwer Academic Publishers, Dordrecht.

Pashigian, B.P. (1985). "Environmental Regulation: Whose Self-Interests Are Being Protected?". Economic Inquiry 23, 551.

Peltzman, S. (1976). "Toward a More General Theory of Regulation". Journal of Law and Economics 19, 211.

Pezzey, J.C.V. (1992). "Sustainable Development Concepts: An Economic Analysis". World Bank Environment Paper No. 2, World Bank, Washington, D.C.

Pezzey, J.C.V., Toman, M.A. (2001). "Progress and problems in the economics of sustainability". In: Tietenberg, T.H., Folmer, H. (Eds.), International Yearbook of Environmental and Resource Economics 2002/2003. Edward Elgar, Cheltenham.

Pezzey, J.C.V., Toman, M.A. (2002). "The Economics of Sustainability: A Review of Journal Articles". Discussion Paper 02-03, Resources for the Future, Washington, D.C.

Phaneuf, D.J., Smith, V.K. (2005). "Recreation Demand Models". In: Mäler, K., Vincent, J. (Eds.), Handbook of Environmental Economics. North Holland/Elsevier Science, Amsterdam.

Pigou, A.C. (1920). The Economics of Welfare. Macmillan, London.

Pildes, R.H., Sunstein, C.R. (1995). "Reinventing the Regulatory State". University of Chicago Law Review 62, 1–129.

Pitchford, R. (1995). "How Liable Should a Lender Be? The Case of Judgement-Proof Firms and Environmental Risk". American Economic Review 85, 1171–1186.

Pizer, W.A. (2002). "Combining Price and Quantity Controls to Mitigate Global Climate Change". Journal of Public Economics 85 (3), 409–434.

Polinsky, A.M. (1971). “Probabilistic Compensation Criteria”. Quarterly Journal of Economics 86 (3), 407–425.

Polinsky, A.M. (1980). “The Efficiency of Paying Compensation in the Pigovian Solution to Externality Problems”. Journal of Environmental Economics and Management 7 (2), 142–148.

Popp, D. (2002). “Induced Innovation and Energy Prices”. American Economic Review 92 (1), 160–180.

Porter, M.E., van der Linde, C. (1995). “Toward a New Conception of the Environment-Competitiveness Relationship”. Journal of Economic Perspectives 9, 97–118.

Portney, P.R. (1990). “Air Pollution Policy”. In: Portney, P.R. (Ed.), Public Policies for Environmental Protection. Resources for the Future, Washington, D.C.

Portney, P.R. (1994). “The Contingent Valuation Debate: Why Economists Should Care”. Journal of Economic Perspectives 8 (4), 3–17.

Portney, P.R., Stavins, R.N. (1994). “Regulatory Review of Environmental Policy: The Potential Role of Health-Health Analysis”. Journal of Risk and Uncertainty 8 (1), 111–122.

Portney, P.R., Weyant, J. (Eds.) (1999). Discounting and Intergenerational Equity. Resources for the Future, Washington, D.C.

Randall, A., Stoll, J.R. (1980). “Consumer’s Surplus in Commodity Space”. American Economic Review 70 (3), 449–455.

Rascoff, S.J., Revesz, R.L. (2002). “The Biases of Risk Tradeoff Analysis: Toward Parity in Environmental and Health-and-Safety Regulation”. The University of Chicago Law Review 69 (4), 1763–1836.

Rasmusen, E., Zupan, M. (1991). “Extending the Economic Theory of Regulation to the Form of Policy”. Public Choice 72 (2–3), 275–296.

Rawls, J. (1971). A Theory of Justice. Harvard University Press, Cambridge, MA.

Reinhardt, F.L. (2000). Down to Earth: Applying Business Principles to Environmental Management. Harvard Business School Press, Boston, MA.

Repetto, R., Dower, R., Jenkins, R., Geoghegan, J. (1992). Green Fees: How a Tax Shift Can Work for the Environment and the Economy. World Resources Institute, Washington, D.C.

Revesz, R.L. (1992). “Rehabilitating Interstate Competition: Rethinking the ‘Race to the Bottom’ Rationale for Federal Environmental Regulation”. New York University Law Review 67, 1210–1254.

Revesz, R.L. (1996). “Federalism and Interstate Environmental Externalities”. University of Pennsylvania Law Review 144, 2341–2416.

Revesz, R.L. (1997a). “Environmental Regulation, Ideology, and the D.C. Circuit”. Virginia Law Review 83, 1717.

Revesz, R.L. (1997b). “The Race to the Bottom and Federal Environmental Regulation: A Response to Critics”. Minnesota Law Review 82, 535.

Revesz, R.L. (1997c). Foundations in Environmental Law and Policy. Oxford University Press, New York.

Revesz, R.L. (1999). “Environmental Regulation, Cost-Benefit Analysis, and the Discounting of Human Lives”. Columbia Law Review 99 (4), 941–1017.

Revesz, R.L. (2001a). “Federalism and Environmental Regulation: A Public Choice Analysis”. Harvard Law Review 115, 553–641.

Revesz, R.L. (2001b). “Federalism and Regulation: Some Generalizations”. In: Esty, D., Geradin, D. (Eds.), Regulatory Competition and Economic Integration: Comparative Perspectives. Oxford University Press, New York, NY.

Rico, R. (1995). “The U.S. Allowance Trading System for Sulfur Dioxide: An Update of Market Experience”. Environmental and Resource Economics 5 (2), 115–129.

Ridker, R.G. (1967). Economic Costs of Air Pollution: Studies in Measurement. Praeger, New York.

Roback, J. (1982). “Wages, rents, and the quality of life”. Journal of Political Economy 90, 1257–1278.

Roberts, M.J., Spence, M. (1976). “Effluent Charges and Licenses Under Uncertainty”. Journal of Public Economics 5, 193–197.

Rose, K. (1997). “Implementing an Emissions Trading Program in an Economically Regulated Industry: Lessons from the SO2 Trading Program”. In: Kosobud, R.F., Zimmerman, J.M. (Eds.), Market Based Approaches to Environmental Policy: Regulatory Innovations to the Fore. Van Nostrand/Reinhold, New York.

Rosen, S. (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition". Journal of Political Economy 82 (1), 34–35.

Rosen, H. (1979). "Wage-Based Indexes of Urban Quality of Life". In: Mieszkowski, P., Straszheim, M. (Eds.), Current Issues in Urban Economics. Johns Hopkins University Press, Baltimore, MA.

Rosenberger, R.S., Loomis, J.B. (2003). "Benefit Transfer". In: Champ, P.A., Boyle, K.J., Brown, T.C. (Eds.), A Primer on Nonmarket Valuation. Kluwer Academic Publishers, Dordrecht.

Rothenberg, J. (1993). "Economic Perspectives on Time Comparisons: Evaluation of Time Discounting". In: Nazli, C. (Ed.), Global Accord: Environmental Challenges and International Responses. MIT Press, Cambridge, MA.

Rowlatt, P., Spackman, M., Jones, S., Jones-Lee, M., Loomes G. (1998). "Valuation of Deaths from Air Pollution". Report prepared by National Economic Research Associates for the Department of Environment, Transport, and the Regions and the Department of Trade and Industry.

Rubin, J.D. (1996). "A Model of Intertemporal Emission Trading, Banking, and Borrowing". Journal of Environmental Economics and Management 31 (3), 269–286.

Sagoff, M. (1993). "Environmental Economics: An Epitaph". Resources (Spring), 2–7.

Sandel, M.J. (1997). "It's Immoral to Buy the Right to Pollute". New York Times (December 15), A29.

Sandel, M.J. (1998). "What Money Can't Buy: The Moral Limits of Markets". The Tanner Lectures on Human Values. Brasenose College, Oxford.

Sarnoff, J.D. (1997). "The Continuing Imperative (But Only from a National Perspective) for Federal Environmental Protection". Duke Environmental Law and Policy Forum 7, 225–319.

Scanlon, T.M. (1991). "The Moral Basis of Interpersonal Comparisons". In: Elster, J., Roemer, J.E. (Eds.), Interpersonal Comparisons of Well-Being. Cambridge University Press, Cambridge, England.

Schelling, T.C. (1995). "Intergenerational Discounting". Energy Policy 23 (4–5), 395–401.

Scheraga, J.D., Sussman, F.G. (1998). "Discounting and Environmental Management". In: Tietenberg, T.H., Folmer, H. (Eds.), International Yearbook of Environmental and Resource Economics 1998/1999. Edward Elgar, Cheltenham, U.K.

Schmalensee, R. (1998). "Greenhouse Policy Architectures and Institutions". In: Nordhaus, W.D. (Ed.), Economics and Policy Issues in Climate Change. Resources for the Future, Washington, D.C.

Schmalensee, R., Joskow, P.L., Ellerman, A.D., Montero, J.P., Bailey, E.M. (1998). "An Interim Evaluation of Sulfur Dioxide Emissions Trading". Journal of Economic Perspectives 12 (3), 53–68.

Seip, K., Strand, J. (1992). "Willingness to pay for environmental goods in Norway: A contingent valuation study with real payments". Environmental and Resource Economics 2, 91–106.

Sen, A. (2000). "The Discipline of Cost-Benefit Analysis". Journal of Legal Studies 29 (2), 931–952.

Sen, A.K. (1970). Collective Choice and Social Welfare. Holden-Day, San Francisco.

Shavell, S. (1981). "A Note of Efficiency vs. Distributional Equity in Legal Rulemaking: Should Distributional Equity Matter Given Optimal Income Taxation?". American Economic Review 71 (2), 414–418.

Shavell, S. (1984). "Liability for Harm Versus Regulation of Safety". Journal of Legal Studies 13, 357.

Shavell, S. (1987). Economic Analysis of Accident Law. Harvard University Press, Cambridge.

Shavell, S. (2005). "Minimum Asset Requirements and Compulsory Liability Insurance as Solutions to the Judgment-Proof Problem". Rand Journal of Economics 36 (1), 63–77.

Shavell, S., Stavins, R.N., Gaines, S., Maskin, E. (1997). "Emissions Trading Will Lead to Less Pollution; What's 'Immoral'?". Letters to the Editor, New York Times (December 17), A30.

Shefrin, H.M., Thaler, R.H. (1988). "The Behavioral Life-cycle Hypothesis". Economic Inquiry 26 (4), 609–643.

Shelby, M., Shackleton, M., Shealy, M., Cristofaro, A. (1997). The Climate Change Implications of Eliminating U.S. Energy (and Related) Subsidies. U.S. Environmental Protection Agency, Washington, D.C.

Shepard, D.S., Zeckhauser, R.J. (1982). "Life-Cycle Consumption and Willingness to Pay for Increased Survival". In: Jones-Lee, M.W. (Ed.), The Value of Life and Safety. North-Holland, Amsterdam, pp. 95–141.

Shepsle, K.A., Weingast, B.R. (1984). "Political Solutions to Market Problems". American Political Science Review 78, 417–434.

Shogren, J.F. (1998). "A Political Economy in an Ecological Web". Environmental and Resource Economics 11 (3–4), 557–570.

Shogren, J.F., Crocker, T.D. (1991). "Risk, Self-Protection, and Ex Ante Economic Value". Journal of Environmental Economics and Management 20 (1), 1–15.

Shogren, J.F., Crocker, T.D. (1999). "Risk and Its Consequences". Journal of Environmental Economics and Management 37 (1), 44–51.

Skumatz, L.A. (1990). "Volume-Based Rates in Solid Waste: Seattle's Experience". Report for the Seattle Solid Waste Utility, Seattle, WA.

Slovic, P. (1987). "Perception of Risk". Science 30 (4), 423–439.

Smith, V.K. (1987). "Nonuse Values in Benefit Cost Analysis". Southern Economic Journal 54 (1), 19–26.

Smith, V.K., Banzhaf, H.S. (2004). "A Diagrammatic Exposition of Weak Complementarity and the Willig Condition". American Journal of Agricultural Economics 86 (2), 455–466.

Snyder, L.D., Stavins, R.N., Wagner, A.F. (2003). "Private Options to Use Public Goods: Exploiting Revealed Preferences to Estimate Environmental Benefits". Faculty Research Working Paper Series, RWP03-013, John F. Kennedy School of Government, Harvard University, Cambridge, MA.

Solow, R. (1981). "Replies to 'Cost-Benefit Analysis: An Ethical Critique' ". AEI Journal on Government and Society Regulation (March/April), 40–41.

Spofford, W.O. Jr. (1984). "Efficiency Properties of Alternative Source Control Policies for Meeting Ambient Air Quality Standards: An Empirical Application to the Lower Delaware Valley". Discussion Paper D-118, Resources for the Future, Washington, D.C.

Stavins, R.N. (1983). Trading Conservation Investments for Water. Environmental Defense Fund, Berkeley, CA.

Stavins, R.N. (1995). "Transaction Costs and Tradeable Permits". Journal of Environmental Economics and Management 29, 133–146.

Stavins, R.N. (1996). "Correlated Uncertainty and Policy Instrument Choice". Journal of Environmental Economics and Management 30, 218–225.

Stavins, R.N. (1997). "Policy Instruments for Climate Change: How Can National Governments Address a Global Problem?". The University of Chicago Legal Forum, 293–329.

Stavins, R.N. (1998). "What Have We Learned from the Grand Policy Experiment: Lessons from SO2 Allowance Trading". Journal of Economic Perspectives 12 (3), 69–88.

Stavins, R.N. (2003). "Experience with Market-Based Environmental Policy Instruments". In: Mäler, K.G., Vincent, J. (Eds.), The Handbook of Environmental Economics, vol. I. North Holland/Elsevier Science, Amsterdam.

Stavins, R.N., Wagner, A.F., Wagner, G. (2003). "Interpreting sustainability in economic terms: dynamic efficiency plus intergenerational equity". Economics Letters 79, 339–343.

Stephenson, K., Norris, P., Shabman, L. (1998). "Watershed-based Effluent Trading: The Nonpoint Source Challenge". Contemporary Economic Policy 16, 412–421.

Stern, A.C. (1982). "History of Air Pollution Legislation in the United States". Journal of Air Pollution Control Association 32, 44.

Stevens, B.J. (1978). "Scale, Market Structure, and the Cost of Refuse Collection". The Review of Economics and Statistics 40, 438–448.

Stigler, G.J. (1971). "The Theory of Economic Regulation". Bell Journal of Economics 2 (1), 3–21.

Stokey, E., Zeckhauser, R. (1978). A Primer for Policy Analysis. W.W. Norton and Company, New York.

Sunstein, C.R. (1996). "Congress, Constitutional Moments, and the Cost-Benefit State". Stanford Law Review 48, 247–309.

Sunstein, C.R. (2002). The Cost-Benefit State: The Future of Regulatory Protection. American Bar Association, Chicago, IL.

Swire, P.P. (1996). "The Race to Laxity and the Race to Undesirability: Explaining Failures in Competition Among Jurisdictions in Environmental Law". Yale Journal on Regulation 14, 67–108.

Taylor, L.O. (2003). "The Hedonic Method". In: Champ, P.A., Boyle, K.J., Brown, T.C. (Eds.), A Primer on Nonmarket Valuation. Kluwer Academic Publishers, Dordrecht.

Tengs, T.O., Adams, M.E., Pliskin, J.S., Safran, D.G., Siegel, J.E., Weinstein, M.C., Graham, J.D. (1995). "Five Hundred Life-Saving Interventions and Their Cost-Effectiveness". Risk Analysis 15, 369.

Tiemann, M. (1999). “Safe Drinking Water Act Amendments of 1996: Overview of P.L. 104-182”. Congressional Research Service Report for Congress 96-722, updated February 8, 1999, Congressional Research Service, Washington, D.C.

Tietenberg, T.H. (1980). “Transferable Discharge Permits and the Control of Stationary Source Air Pollution: A Survey and Synthesis”. Land Economics 56, 391–416.

Tietenberg, T.H. (1985). Emissions Trading: An Exercise in Reforming Pollution Policy. Resources for the Future, Washington, D.C.

Tietenberg, T.H. (1995). “Tradeable Permits for Pollution Control When Emission Location Matters: What Have We Learned?”. Environmental and Resource Economics 5 (2), 95–113.

Tietenberg, T.H. (1997a). “Tradeable Permits and the Control of Air Pollution in the United States”. Paper prepared for the 10th Anniversary Jubilee edition of Zeitschrift Fürangewandte Umweltforschung.

Tietenberg, T.H. (1997b). “Information Strategies for Pollution Control”. Paper presented at the Eight Annual Conference, European Association of Environmental and Resource Economists, Tilburg, the Netherlands.

Tietenberg, T.H. (2003). “Tradeable Permits in Principle and Practice”. Paper prepared for presentation at “Twenty Years of Market-Based Instruments for Environmental Protection: Has the Promise Been Realized?” Donald Bren School of Environmental Science and Management, University of California, Santa Barbara, August 23–24, 2003, Colby College.

Tolley, G.S., Kenkel, D., Fabian, R. (1994). Valuing Health for Policy: An Economic Approach. University of Chicago Press, Chicago, IL.

Tschirhart, J.T. (1984). “Transferable discharge permits and the control of stationary source air pollution: A survey and synthesis”. In: Crocker, T.D. (Ed.), Economic Perspectives on Acid Deposition Control. Butterworth, Boston.

U.S. Congress, Office of Technology Assessment (1992). Building Energy Efficiency. U.S. Governmental Printing Office, Washington, D.C.

U.S. Council on Environmental Quality (1997). Environmental Justice. Guidance under the National Environmental Policy Act. U.S. EPA, Washington, D.C.

U.S. Environmental Protection Agency (1990). “Environmental Investments: The Cost of a Clean Environment”. Report of the Administrator to Congress, U.S. EPA, Washington, D.C.

U.S. Environmental Protection Agency (1991). “Economic Incentives, Options for Environmental Protection”, Document P-2001, U.S. EPA, Washington, D.C.

U.S. Environmental Protection Agency (1992). “The United States Experience with Economic Incentives to Control Environmental Pollution”. EPA-230-R-92-001, U.S. EPA, Washington, D.C.

U.S. Environmental Protection Agency (1993). “Benefits Transfer: Procedures, Problems and Research Needs”. Report prepared by the Office of Policy, Planning and Evaluation, EPA/230/R-93/018, U.S. EPA, Washington, D.C.

U.S. Environmental Protection Agency (1999). The Benefits and Costs of the Clean Air Act, 1990 to 2010, Prepared for the U.S. Congress, U.S. Environmental Protection Agency, Washington, D.C.

U.S. Environmental Protection Agency (2000a). Guidelines for Preparing Economic Analyses. U.S. EPA, Washington, D.C.

U.S. Environmental Protection Agency (2000b). Handbook for Non-Cancer Health Effects Valuation. Science Policy Council, Social Sciences Discussion Group, U.S. EPA, Washington, D.C.

U.S. Environmental Protection Agency (2001). The United StatesExperience with Economic Incentives for Protecting the Environment. U.S. EPA, Washington, D.C.

U.S. Environmental Protection Agency (2002). Cost of Illness Handbook. U.S. EPA, Washington, D.C.

U.S. Environmental Protection Agency, Office of Policy Analysis (1985). Costs and Benefits of Reducing Lead in Gasoline: Final Regulatory Impact Analysis. U.S. EPA, Washington, D.C.

U.S. General Accounting Office (1992). Toxic Chemicals: EPA’s ToxicsRelease Inventory is Useful But Could be Improved. U.S. GAO, Washington, D.C.

U.S. Office of Management and Budget (2003). Circular A-4, Regulatory Analysis, Washington, D.C.

Viscusi, K. (1992). Fatal Trade Offs: Public and Private Responsibilities for Risk. Oxford University Press, New York.

Viscusi, K. (1993). "The Value of Risks to Life and Health". Journal of Economic Literature 31 (4), 1912–1946.

Viscusi, W.K., Aldy, J.E. (2003). "The Value of a Statistical Life: A Critical Review of Market Estimates throughout the World". Journal of Risk and Uncertainty 27 (1), 5–76.

Viscusi, K., Magat, W., Huber, J. (1991). "Pricing environmental health risks: A survey assessment of risk-risk and risk-dollar trade-offs for chronic bronchitis". Journal of Environmental Economics and Management 21, 32–51.

Wahl, R.W. (1989). Markets for Federal Water: Subsidies, Property Rights, and the Bureau of Reclamation. Resources for the Future, Washington, D.C.

Waugh, F.V. (1928). "Quality Factors Influencing Vegetable Prices". Journal of Farm Economics 10 (2), 185–196.

Weitzman, M.L. (1974). "Prices vs. Quantities". Review of Economic Studies 41 (4), 477–491.

Weitzman, M.L. (1978). "Optimal Rewards for Economic Regulation". American Economic Review 68 (4), 683–691.

Weitzman, M.L. (1994). "On the Environmental Discount Rate". Journal of Environmental Economics and Management 26 (2), 200–209.

Weitzman, M.L. (1998). "Why the Far-Distant Future Should Be Discounted at Its Lowest Possible Rate". Journal of Environmental Economics and Management 36 (3), 201–208.

Weitzman, M.L. (2003). Income, Capital, and the Maximum Principle. Harvard University Press, Cambridge, MA.

Wertz, K.L. (1976). "Economic Factors Influencing Households' Production of Refuse". Journal of Environmental Economics and Management 2, 263–272.

Wildavsky, A. (1980). "Richer is Safer". The Public Interest 60, 27–29.

Willig, R. (1976). "Consumer's Surplus without Apology". American Economic Review 66 (4), 589–597.

World Commission on Environment and Development (1987). "Our Common Future". Oxford University Press, Oxford.

Zerbe, R.O. (1970). "Theoretical Efficiency in Pollution Control". Western Economic Journal 8, 364–376.

This page intentionally left blank

Chapter 9

REGULATION OF HEALTH, SAFETY, AND ENVIRONMENTAL RISKS

W. KIP VISCUSI

Department of Economics, Owen Graduate School of Management, and School of Law,
Vanderbilt University, and National Bureau of Economic Research

Contents

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)01009-2

Abstract

This paper provides a systematic review of the economic analysis of health, safety, and environmental regulations. Although the market failures that give rise to a rationale for intervention are well known, not all market failures imply that market risk levels are too great. Hazard warnings policies often can address informational failures. Some market failures may be exacerbated by government policies, particularly those embodying conservative risk assessment practices. Labor market estimates of the value of statistical life provide a useful reference point for the efficient risk tradeoffs for government regulation. Guided by restrictive legislative mandates, regulatory policies often strike a quite different balance with an inordinately high cost per life saved. The risk-risk analysis methodology enables analysts to assess the net safety implications of policy efforts. Inadequate regulatory enforcement and behavioral responses to regulation may limit their effectiveness, while rising societal wealth will continue to generate greater levels of health and safety.

Keywords

risk, regulation, value of statistical life, health and safety, environment, risk–risk analysis, benefit-cost analysis, cost effectiveness

JEL classification: D81, L51, K2, Q50

1. The rise of risk regulation

Beginning in the 1970s, there was a major wave of health, safety, and environmental regulation. In the United States, the government established new regulatory agencies with broad regulatory responsibilities for risk and environmental policy. Some of these agencies addressed risks to people, such as mortality and morbidity risks, while others focused on dangers to natural resources, such as endangered species, which affect people indirectly. Among these new federal agencies were the Environmental Protection Agency (EPA), the Occupational Safety and Health Administration (OSHA), the National Highway Traffic Safety Administration (NHTSA), the Nuclear Regulatory Commission (NRC), and the Consumer Product Safety Commission (CPSC). In some cases, these agencies inherited functions that already existed but had been scattered among other regulatory agencies, such as the U.S. Department of Interior. In other instances, these agencies had entirely new legislative mandates that greatly expanded the scope and character of regulation.

The creation of these agencies did not take place because there was a sudden increase in societal risk levels. Indeed, individual mortality risks levels of almost all kinds had been declining throughout the twentieth century and had been continuing to decline. For example, there were declines in accident rates at home and occupational fatalities throughout the past century, long before there was any government regulation in place.[1] One exception to this mortality risk trend was the automobile fatality rate, which had increased overall due to the greater usage of motor vehicles. The motor vehicle death rate per 100,000 population peaked in 1937 and has exhibited an uneven but downward trend thereafter. However, the fatality rate per mile driven had continued to decline quite steadily so that in 2003 it was 1.56 per 100,000 vehicle miles, as compared to 4.88 per 100,000 vehicle miles in 1970 and 14.68 per 100,000 vehicle miles in 1937. Auto safety became the first target of the emerging consumer movement that had been fostered in part by the auto safety efforts of Ralph Nader (1965). Environmental risks displayed a somewhat different temporal trajectory, as many concentrations of air and water pollutants increased until the early 1970s and declined thereafter once the environmental regulations were instituted, as documented by Freeman (2002). These regulations in turn have led to steady improvements in environmental quality with only a few exceptions, such as carbon dioxide, which has continued to rise. Support for both consumer protection and risk regulation generally also can be traced to rising societal wealth levels that increased individual valuations of risk reduction, thus increasing the demand for regulatory intervention.

The role of the courts with respect to risk has also increased since the 1970s. A noteworthy period was the mid-1980s when liability insurance premiums for general liability and medical malpractice approximately tripled in a two-year period, prompting claims that there was a "liability crisis." The counterargument to the crisis view was that

[1] The statistics cited in this paragraph are drawn from the National Safety Council (2004).

the upsurge in premium rates reflected the normal ebb and flow in the insurance industry known as the underwriting cycle. Periodic declines in interest rates led to an associated increase in premium rates because the return insurers received on the invested premium amounts had declined. Although the increases in liability costs abated, the costs imposed by tort liability remained at or above the mid-1980s levels, in part because the scope of liability rules had expanded.

The expanded role of liability in some cases encouraged additional regulatory activity. A notable example is that of asbestos. Traditionally there had been comparatively lax regulation of asbestos, in that the risks were substantial and could be reduced at modest cost. The wave of asbestos-related lawsuits spurred the imposition of stringent occupational and environmental standards. After a period of inadequate regulation of asbestos, both EPA and OSHA imposed a succession of asbestos regulation standards that greatly altered the tradeoff between cost and risk and, in the view of some observers, such as Breyer (1993), erred on the side of being too stringent.

This pattern of liability spurring further regulation has continued in other contexts as well. Three principal examples of the interaction of regulation and litigation are breast implants, lead paint, and cigarettes. In the case of breast implants, these medical devices had been exempted from serious scrutiny from a safety standpoint until a series of successful lawsuits led the FDA to suspend the use of silicone breast implants until the safety issues could be resolved, as described in Hersch (2002). Lutter and Mader (2002) found that for lead paint the causality was in the other direction as 1978 CPSC regulations had banned the use of lead paint, but because the paint remains in older dwellings there have been a spate of lead paint-related lawsuits, including major class actions. For the past decade there has been a synergistic relationship between regulation and litigation against the cigarette industry, much of which was stimulated by the settlement of a series of state attorney general lawsuits against the industry for close to $250 billion, as discussed in Viscusi (2002).

While the interactive effects of regulation and litigation have generated substantial controversy, there is a general consensus that there should be health, safety, and environmental regulation. The main debate is over the targets for regulatory action, the modes of intervention selected, and the stringency of the regulation. A substantial body of economic research has focused on striking the appropriate balance between risk reduction and cost and on ways in which regulations can be designed to maximize social welfare.[2] While this formulation has widespread acceptance in the economics literature, economists' concern with maximizing social welfare often is inconsistent with more narrowly defined approaches taken by regulatory agencies, which may focus on less balanced objectives, such as the promotion of clean air irrespective of other concerns, such as cost.

[2] Among the more comprehensive treatments of this issue are Breyer (1993), Viscusi (1992), and Sunstein (2002).

The focus of this chapter is on the theoretical and empirical methodologies that have been developed to analyze the substantive economic issues in this area. In some instances, the economic frameworks can be used to provide guidance for policy design, while in other cases the emphasis is on assessments of policy performance for what continues to be the most costly area of government regulation. Ideally, if economic efficiency is our objective, these efforts should also produce commensurate benefits so that these policies will benefit society on balance. Moreover, policies should strive to produce the greatest spread between benefits and costs.

The chapter is organized as follows. In section 2, I construct the basic individual choice model involving a risky decision. Even if choices are not ideal and there is a rationale for intervention, how people would choose to bear risks in an efficient context frequently serves as the reference point for guiding policy interventions. After I examine the sources of market failure in section 3, section 4 develops the analysis of informational types of regulatory interventions, which address the frequent shortcoming in risk contexts. Before considering the use of benefit-cost analysis in section 7, this chapter first explores risk assessment procedures in section 5 and risk valuation in section 6. How the newly developed risk-risk analysis approach would alter these policy judgments is the subject of section 8, while section 9 examines enforcement of risk and environmental regulations.

2. Model of risky decisions

Before considering the role of government regulation and tort liability, it is useful to begin with a model of rational individual choice. The effect of inadequacies in such choices on market performance is a principal rationale for intervention. In addition, the thought experiment of how people would make decisions involving risk if markets functioned perfectly and choices were fully rational provides a useful reference point for establishing the appropriate tradeoffs between risk and cost. Thus, what balance would individuals themselves choose to strike between these two components of decisions?

For concreteness, let us focus on two decisions—a private decision to invest in good health as well as a market decision involving the individual's choice of a potentially hazardous job. Within the job choice context, workers will require additional wage premiums to face additional risk. The model could be developed in parallel fashion for product risks rather than job risks. The product risk counterpart is directly analogous. Just as workers will require higher wages to work on jobs posing greater risk, consumers will require a reduced price to lead them to buy more dangerous products. Since most empirical studies of money-risk tradeoffs have used labor market data on wages and risk to impute these tradeoffs, the risky job choice decision will be the case study examined below.

Note that in both the product market case and job market case, the parties are in a contractual relationship. A company with hazardous jobs posing risks that are known to workers will not attract individuals to the jobs unless the company pays a sufficient wage to match the workers' utility level in alternative employment. In contrast, many risky decisions, such as the decision to pollute, involve risks to third parties that are not part of a contractual relationship. These third party risks will be addressed further within the discussion of externalities.

Suppose that the individual can choose the level of job safety s from a range of continuous market opportunities $w(s)$, where $w_s > 0$ and $w_{ss} \leqslant 0$. The individual risk level is also dependent on the level of health-enhancing expenditures, h. For concreteness let there be two states of the world, good health and death, for which the pertinent utility functions are u and v, where v is a bequest function. Assume that people are risk-averse or risk-neutral and that being alive is preferable to the alternative: $u(x) > v(x) > 0$, $u'(x) > v'(x)$, and $u''(x)$, $v''(x) \leqslant 0$. The probability π of being in the good health state $\pi(s,h)$ increases with the safety level s and the health-related expenditures, h. The risk of death $1-\pi(s,h)$ is assumed to be perceived accurately. Finally, let y be the individual assets.

An expected utility (EU) maximizer will consequently choose s and h to

$$\operatorname*{Max}_{s,h} EU = \pi(s,h)u((y+w(s)-h)) + (1-\pi(s,h))v(y+w(s)-h). \tag{1}$$

The first-order condition for the optimal value of h is

$$\frac{1}{\pi_h} = \frac{u-v}{\pi u' + (1-\pi)v'}, \tag{2}$$

and it can be shown that the comparable requirement for s is

$$\frac{-w_s}{\pi_s} = \frac{u-v}{\pi u' + (1-\pi)v'}. \tag{3}$$

The term on the left side of equation (3) is the negative of the marginal change in wages in response to increases in the safety level s divided by the marginal effect of the job safety level s on the probability of survival. If the safety level metric s is equivalent to the probability of survival, then $\pi_s = 1$. For small changes in s, the value of $-w_s$ will then equal the marginal wage increase needed to face added risk and, equivalently, the negative of the marginal wage increase for greater levels of safety. This compensating differential term is set equal to the utility level difference between the two states, which is normalized by dividing by the expected marginal utility of consumption.

In the case of equation (2), the individual sets the marginal value of risks to life as revealed by health expenditures equal to the difference in utility level between the two states divided by the expected marginal utility of consumption. For the job safety choice reflected in equation (3), the worker also sets the marginal value of risks to life equal to the difference between the utility levels in the two states divided by the expected marginal utility of consumption. Putting the two conditions above together, we have

$$\frac{1}{\pi_h} = \frac{-w_s}{\pi_s} = \frac{u-v}{\pi u' + (1-\pi)v'} = \text{VSL}, \tag{4}$$

where VSL is the marginal value of statistical life, which is reflected throughout one's risk-taking and risk-reducing decisions.

These results have several implications for policy. First, if people are cognizant of the risks, expected utility maximizers will choose the risk-reducing and risk-increasing activities so as to reflect a common VSL, which is a measure of their competing effects on individual welfare. The efficient level of risk is not zero in general. Second, this balancing will yield identical tradeoff rates across different domains of choice if there are smooth and continuous choices available. Whether the choice is a risky job, a risky product, a dangerous neighborhood, or beneficial health expenditures, if there is a continuum of choices people will set the tradeoff rates between risk and money to be the same in these different arenas. If particular areas of expenditure, such as allocations for regulations, are out of line with these tradeoff rates, it serves as a signal that economic efficiency could be enhanced.

Finally, as will be discussed below, equation (4) characterizes the value of statistical life as revealed through labor market decisions. Considerable empirical work has been devoted to estimating VSL levels in the labor market and in other contexts as well, where these values in turn will provide a yardstick for assessing the tradeoffs being made by regulatory policy. For example, if the tradeoff rate between risk and money is too low in a particular area compared to the tradeoffs the person is willing to make in other contexts, one might explore whether such highly cost-effective efforts can be expanded. These values in turn have served as the unit benefit values in assessing the benefits of government regulation.

The use of labor market VSL estimates in other policy contexts raises a variety of interesting theoretical concerns that have come under the general heading of the benefits transfer problem. First, the mix of people affected by a policy may have different attitudes toward incurring risk due to the heterogeneity of risk tradeoffs. Second, even for a particular utility function structure, the estimated VSL will decrease with the size of the baseline risk and with the extent of the risk reduction for which the willingness to pay value is being elicited. Third, labor market estimates reflect willingness to accept values, whereas policy evaluation hinges on willingness to pay values. For small changes in risk these values should be the same, but often are different in experimental studies, for which the willingness to accept values may be much larger.

The development here has been in terms of multiple risk-related activities that jointly affect the probability of survival π. The subsequent sections will frequently simplify the formulation of the probabilistic component by focusing on the decision context in which there is only a single risky activity rather than multiple choices that influence risk. With a more restricted model of that type, the person chooses the overall fatality risk directly within the context of a single choice situation, such as the riskiness of one's job. Adopting simplifications along these lines is often instructive, but at times it will also prove useful to recall the more general result that optimal risk taking will involve comparable risk-money tradeoffs across different domains of choice.

3. Sources of market failure

If markets functioned perfectly then there would be no need for government risk regulation. People would knowingly choose products, jobs, and activities to maximize their expected utility, and outcomes would be efficient. The various departures from the idealized economic world consist of a series of possible market failures to be reviewed briefly below. Increasingly, regulatory agencies are beginning to discuss these market failures in their efforts to justify intervention rather than simply assuming that new regulations are always desirable. Even if there is a shortcoming of markets, whether a regulation will in fact enhance societal welfare must be assessed as well. Thus, the existence of a market imperfection does not necessarily imply that a particular regulation will be beneficial.

3.1. Imperfect perception of risk

3.1.1. Biases in risk beliefs

A potential shortcoming in individual choices involving risk can arise if people don't understand the risks involved so that there is a systematic bias in their risk beliefs. It is rare that people know the exact risks posed by their risky decisions. What is the reduced probability of death that results from purchasing a car with side curtain air bags or buying a house in a less polluted neighborhood? People may not have precise and accurate assessments of these or many other risks they face. However, a lack of such knowledge does not necessarily imply that there is a market failure that would give regulation or tort liability a role to play. We must first examine the nature and extent of the problem and how it affects decisions as compared to choices people would make with an accurate information reference point. Risk underestimation leads to too much risky activity, and risk overestimation leads to the opposite problem.

In many instances, it does not matter a great deal if risk perceptions are accurate or that if we make sound choices with respect to risk. The stakes are often small and involve few costs of error that cannot be corrected with better future decisions. Errors may be critically important with respect to catastrophic events, which tend to have small probabilities. Catastrophic risks to individuals and property are the main target of both regulatory policy and tort liability.[3]

For concreteness, consider the product market case in which there is a risk p^* of injury from the product. In terms of the previous multi-risk model, p^* corresponds to the value of $1 - \pi(s, h)$. Suppose the consumer has a subjective assessment of this risk given by p, with associated precision γ, where γ is the equivalent number of draws from a Bernoulli urn that is reflected in the person's subjective beliefs.[4] Higher values of γ imply more precise probabilistic beliefs.

[3] A useful survey of issues pertaining to low probability-high consequence events is that by Camerer and Kunreuther (1989). A further discussion of catastrophic risks appears in Posner (2004).

[4] The discussion in this chapter assumes that probabilistic beliefs can be characterized by a beta distribution with the parameterization to be discussed in the text.

Rational expected utility maximizers engaged in some single period decision should treat subjective risk beliefs as equivalent to objective risk measures. Thus, if $p = p^*$, the subsequent choices will be expected utility maximizing, based on the true risk levels. If $p < p^*$, people underestimate the risk, while if $p > p^*$ people overestimate the risk. The fact that people have subjective risk beliefs and may not be fully informed consequently does not imply that people necessarily underestimate the risk and engage in too much risk taking behavior, since it is not necessarily the case that $p < p^*$.

How risk beliefs diverge from the actual risk levels is an empirical issue. For hidden risks, such as the presence of trace amounts of benzene in Perrier, it may be that people underestimated the risk before the presence of benzene was discovered, since this was a situation of ignorance. People simply did not know that there was benzene in Perrier, though the actual risks were never that great. In situations of ignorance, people will tend to respond inadequately to risks, as has been shown with respect to natural disasters by Kunreuther (1978). Once the risks are publicized, as in the Perrier benzene case, it is likely that the public overestimated the risks. Indeed, even after eliminating the presence of benzene, the product never recovered its earlier share of the bottled water market. Thus, substantial media coverage may lead to risk overestimation rather than accurate risk beliefs.

In some extreme instances, the media coverage of negligible risks may lead regulators to undertake policy actions not warranted by actual risk levels. McClelland, Schulze, and Hurd (1990) found that the public has a very inaccurate perception of the risks posed by hazardous waste sites but that these erroneous beliefs adversely affect housing prices. These and similar results raise the policy question of the extent to which policies should address risks that are perceived and not real, but nevertheless may have significant economic consequences in that markets may react to perceptions rather than actual risks.

While discussions of the role of tort liability and risk regulation often hypothesize that people underestimate risks, that pattern does not necessarily hold. Errors in risk beliefs do not necessarily imply that people always underestimate risks. Biases in risk beliefs are not random but vary in a variety of systematic ways. An important risk dimension that has received extensive empirical study is the magnitude of the risk. In the case of mortality risks, for example, the general pattern is that people overestimate small risks and underestimate large risks.[5] Thus, people tend to overestimate small risks such as those from botulism and lightning strikes, and they underestimate more consequential risks such as the risk of heart disease. To the extent that job risks and product risks are small, people may on balance overestimate these risks more often than underestimating them, provided of course that the risks are not hidden and there is some awareness that a risk is present. Assessments of whether there is a market failure stemming from risk underestimation consequently depends on the level of the risk and the associated level of risk perceptions in the particular market context.

[5] See the early study by Lichtenstein et al. (1978) of this relationship. Fischhoff et al. (1981) examine these and many other salient biases in risk beliefs.

The tendency to overestimate small risks has an interesting corollary as well. If people exaggerate the importance of small risks, then they will overvalue policies that reduce the risk to zero. The reason is that they will perceive a greater reduction in the risk than is actually being generated. Thus, people will be willing to pay a zero risk premium. People will be willing to pay more for a reduction of a risk to zero than they would for a comparable actual risk reduction that does not reduce the risk to zero.

Reducing the risk to zero may offer additional benefits to the risky decision maker. Once the risk is completely eliminated, there is no reason to worry about the risk or factor the presence of a risk into one's decision making. Unless the anxiety reduction benefits of reducing a risk to zero are substantial, it is generally believed that the premium people place on zero risk levels represents a form of irrationality. The importance of this phenomenon has not been lost on government regulators. We are usually assured that our food is "safe" rather than being told that there is a low but nonzero probability that the food will make us ill.

A substantial literature has examined a wide variety of other forms of systematic biases in risk beliefs. For example, people tend to overreact to increases in the risk level from its accustomed risk. Researchers have referred to this phenomenon as reference risk bias and status quo bias.[6] Thus, changes in a product that increase the risk will tend to produce an exaggerated response, implying a higher rate of tradeoff than comparable decreases in risk.

Consequently, people may be willing to pay modest amounts for significant reductions in the risk posed by a product. However, if the risk of the product were to increase by a small amount, they often refuse to purchase the product altogether and, if they are willing to purchase it at all, they require a much more substantial price reduction than the amount they were willing to pay for the risk decrease.[7]

Substantial publicity regarding a risk may lead to risk overestimation as well.[8] This result runs counter to the usual economic assumption that additional information leads to more informed judgments. The reason is that the information often takes the form of publicizing only the number of adverse risk events rather than the risk frequency. We often hear reports of the number of ATV deaths or mad cow disease victims, but these statistics are seldom put in the context of the total number of ATVs in use or the size of the beef-eating population that is at risk. What is being publicized is the numerator of the risk calculation, rather than the numerator and the denominator or the overall risk frequency.[9]

[6] For discussion of these effects, see Viscusi, Magat, and Huber (1987) and Samuelson and Zeckhauser (1988) respectively.

[7] Viscusi, Magat, and Huber (1987) provide supporting empirical evidence for two representative consumer products.

[8] Fischhoff et al. (1981) discuss the role of publicity with respect to risk beliefs.

[9] Viscusi and Zeckhauser (2004) term this phenomenon "the denominator blindness effect," and they document the effect in several contexts.

The extent of the publicity devoted to the number of deaths in the numerator of a risk calculation than the exposed population that comprises the denominator of the risk calculation also may differ depending on the risk context. A dramatic risk event, such as the 9/11 terrorist attack on the World Trade Center, garnered much more media coverage than routine auto accidents even though more people are killed every month in the U.S. on highways than in this worst case terrorist attack. Tornadoes, earthquakes, and hurricanes also generate dramatic media coverage and likely risk overestimation. The role of responses to media coverage makes the extent of publicity given to the hazard a pertinent factor in assessing the direction of bias in risk beliefs.

The substantial publicity given to smoking risks has led to overestimation of these risks compared to scientists' estimates of the actual level of the hazard. For example, U.S. government estimates of the lung cancer risks of smoking over a smoker's lifetime range from 0.06 to 0.13, whereas the lung cancer risk perceptions by the general public range from 0.43 to 0.48 for national surveys undertaken from 1985 to 1998.[10] Similar results have been found for Spain and Taiwan even though the warnings environment in those countries is different.[11]

3.1.2. Risk ambiguity

Not all risks are known with precision. Thus, it is not simply the level of these risk beliefs that may be consequential, but also their precision. Examination of the role of the tightness of risk beliefs comes under the general heading of risk ambiguity.

A potential bias in people's treatment of risk that has played a prominent role in the literature stems from the influence of risk ambiguity. The classic Ellsberg (1961) Paradox focuses on people's attitude with respect to ambiguous risks, that is, risks for which the subjective probability judgment with respect to the probability p represents a situation of imperfect information.[12] Lower values of γ reflect less precise probabilistic beliefs. In situations in which individuals have an opportunity to win a prize, the standard Ellsberg Paradox result is that people would prefer precise probabilities of success to imprecise probabilities, for any given mean value of the probability. Analogous results also hold with respect to the chance of a loss as people exhibit ambiguity aversion with respect to imprecise probabilities of adverse effects. Thus, in situations in which risks are ambiguous, such as the imprecisely understood risks of mad cow disease, one would expect people to exhibit ambiguity aversion and be reluctant to incur these imprecisely understood hazards, as compared to precisely understood risks of the same

[10] Viscusi (2002) presents the series of studies on U.S. smoking behavior.

[11] Estimates for Taiwan are reported by Liu and Hsieh (1995), and findings for Spain are discussed in Viscusi (2002).

[12] Raiffa (1961) provides interesting lottery counterexamples to the Ellsberg Paradox to demonstrate the irrationality of succumbing to the Ellsberg Paradox. The literature on risk ambiguity is quite extensive. See Camerer and Weber (1992) for a survey and Kunreuther, Hogarth, and Meszaros (1993) for an analysis of insurance market effects.

magnitude. The unwillingness to incur ambiguous risks will consequently deter people from making efficient market choices. They incur too small a risk in situations in which the risks are identified but are not well understood.

Because citizen preferences influence pressures for government policies, irrational responses to ambiguous risks have led to similar biases in the formulation of government policy. An interesting case study of the role of ambiguity aversion is government regulation of synthetic risks as opposed to natural carcinogens. Many seemingly safe foods pose some minor carcinogenic risk, including mustard and coffee, and many of these natural carcinogens pose greater risks than some synthetic chemicals that are regulated.[13] Particularly for regulations by the Food and Drug Administration, the magnitude of the cancer risk is not especially influential in determining whether a particular chemical is regulated. What does play a dominant role is whether the chemical is a synthetic.[14] Focusing on this particular character of the risk rather than the magnitude of the hazard is a reflection of how risk ambiguity aversion has moved from the realm of being an irrationality that has been of academic interest to a property of government policy making.

How the lack of full information plays out in a particular market context depends on the character of the risk. Most hazards involve low probability events and risks that are not fully understood. Each of these attributes will lead people to be more reluctant to incur risks than they would be if fully informed of the risk. The opposite problem arises for some very large risks and risks that are truly hidden, such as undisclosed prescription drug interactions. The early and extensive regulation of information disclosures for drugs is reflective of the importance of such disclosures when the risks are hidden.

3.2. Irrational behavior and addiction

While most analyses of departures from efficient market conditions pertaining to risk decisions involve inadequate risk information, there are other possible shortcomings of choices as well. The most prominent of these failures pertains to addiction. From an economic standpoint, addiction generally pertains to market situations in which consumption of a good now leads one to prefer greater levels of consumption of that good in the future. Altering that consumption pattern also imposes substantial costs, making it difficult for the individual to change such consumption behavior. Smoking, drinking, illegal drug use, and overeating have played the greatest roles as case studies in the addiction literature.

[13] For evidence on carcinogenicity of different chemicals and foods, see Ames and Gold (1990) and Ames, Profet, and Gold (1990).

[14] Econometric evidence documenting this bias for the carcinogen potency data base set of chemicals is presented in Viscusi (1998). For a sample of 267 synthetic chemicals, 50 percent were not carcinogenic, which is similar to the 52 percent figure for the 98 natural chemicals studied. However, 47 percent of these synthetic chemicals were regulated by the FDA, compared to 34 percent for the natural chemicals, even though the carcinogenicity of the synthetic chemicals was lower.

Whether there is a market failure with respect to addiction depends on the nature of the choice process. Under the rational addiction models, such as those espoused by Becker and Murphy (1988) and by Becker, Grossman, and Murphy (1994), individuals anticipate their future addiction and rationally choose to become addicted. In their model, fully anticipated addictions do not necessarily involve a market failure. Rather, people could rationally choose to become addicted to a product even though giving up the addiction may be costly and the product itself may be quite dangerous.

The basic rational model of addiction developed by Becker and Murphy (1988) and Becker, Grossman, and Murphy (1994) recognizes that current consumption of the addictive commodity c_t depends on past consumption of that good c_{t-1} in the previous period. Models postulating irrational behavior likewise make that assumption about the intertemporal dependence of consumption. A distinctive feature of rational addiction models is that when people choose to become addicted, they anticipate that current consumption will increase their desired future consumption levels c_{t+1}.

Within an infinite time horizon framework, the rationally addicted consumer will choose consumption of the addictive product to

$$\text{Max} \sum_{t=1}^{\infty} \beta^{t-1} u(c_t, c_{t-1}, y_t, e_t), \tag{5}$$

where r is the rate of interest, β is the discount factor $1/(1+r)$, y_t is income in year t, e_t is unmeasured life cycle variables, and the utility maximization is subject to a lifetime budget constraint. After assuming that the utility function is quadratic in y_t, c_t, and e_t, Becker, Grossman, and Murphy (1994) show that the lifetime consumption trajectory for an addictive good, such as cigarettes, satisfies

$$c_t = \theta c_{t-1} + \beta\theta c_{t+1} + \theta_1 p_t + \theta_2 e_t + \theta_3 e_{t+1}, \tag{6}$$

where p_t is the price of cigarettes in year t, and the θ terms are parameters that depend on the individual's utility function and discount rate. As in standard models of consumer behavior, increasing the current price p_t does have a negative effect on consumption. For addictive goods, past consumption leads to higher desired levels of current consumption.[15] The greater the effect of past consumption on current consumption, as reflected in higher values of θ, the more addictive the commodity is. Similarly, if addiction is rational and anticipated, future price levels that are anticipated should have a greater effect on consumption than unanticipated future price shocks. For analogous reasons, long-run price effects will have a greater effect on consumption than is reflected in the short-run price effect.

A quite different perspective offered by Schelling (1984) is that individuals who are addicted are engaged in a continuing battle for self-control. If people behave myopically and do not place a sufficiently great weight on their future selves, then according to this

[15] This condition also requires that p_t, e_t, e_{t+1}, and the marginal utility of wealth are held constant.

view people will tend to engage in too much behavior that is addictive, leading to a rationale for either direct regulation or taxation of the addictive product.

A similar market failure has been hypothesized by Gruber and Köszegi (2001). In their model, people suffer from time inconsistency in their choices. For example, people may exhibit hyperbolic discounting in which the rate of time preference for immediate rewards is much greater than for deferred payoffs. Moreover, their formulation shares some empirical predictions with the rational addiction framework, making it difficult to disentangle whether the rational addiction model or the time inconsistency model is operative. In particular, the Becker, Grossman, and Murphy (1994) empirical result regarding the role of anticipatory behavior does not rule out the possibility of time inconsistency.

More recently, Bernheim and Rangel (2004) have drawn on the literature from neuroscience and cognitive psychology to develop a model of addictive behavior in which people initially engage in addictive behavior as a result of a mistaken consumption decision. The addicted consumers nevertheless continue with their addictive behavior as environmental cues trigger their subsequent consumption decisions.

The rationale for regulating addictive products often depends on the overall social implications of addiction. Potential harm to the addicted individual is of course an important component. However, addictions may impose other costs as well. Drunk drivers may kill pedestrians and other motorists, and heroin addicts may resort to crime to support their habit. Thus, externalities are often intertwined with most regulatory analyses of addictive behaviors.

3.3. Externalities

The classic rationale for government regulation in the risk and environmental area is the presence of externalities. While environmental pollution is the most common textbook example of an externality, externalities are present elsewhere as well. In workplace contexts, risky behavior on the part of some workers may lead to co-worker injuries, which has prompted government regulation of work practices and workplace conditions. Consumption choices by individuals may affect others adversely as well.

A company generating an externality on a third party will choose the level of activity to maximize its profits, which can also be recast as maximizing the difference between its benefits minus costs. The optimizing polluter will set marginal benefits equal to marginal costs, as in usual optimization frameworks. If the externalities are imposed on third parties, such as the general public, and there is no liability and no regulatory sanctions, then the marginal costs to the company will understate the social marginal costs, and the polluter companies will engage in too much of the risky activity. The existence of liability or regulatory sanctions for such harms will induce greater levels of care by imposing higher marginal costs on the firm. Using this formulation, Cohen (1986) illustrates the role of marginal benefits and marginal costs using oil spills as a case study. Based on reasonable empirical assumptions, he concludes that the Coast Guard's oil spill enforcement efforts appear to generate benefits in excess of costs. Whether the enforcement efforts are in fact socially optimal in terms of maximizing the spread between benefits and costs is less clear.

A notable recent example that has attracted considerable policy attention is with respect to sport utility vehicles, SUVs.[16] These vehicles create two kinds of externalities for other cars. First, because SUVs have greater mass than the average car, they pose a greater threat to other vehicles because the risks posed in a crash involving two vehicles increase with respect to the ratio of the relative weight of the two vehicles. This risk factor is not a distinctive aspect of SUVs, as it is also shared by heavy cars. Second, the structure of SUVs also may be influential as well. Because many SUVs are based on frames adapted from small trucks, some SUVs have stiffer frames than cars and also often have bumper heights that do not align with the height of bumpers of cars, thus generating more damage to the other vehicle than if a car with the same weight had been involved in the crash.

Though there has yet to be any formal regulation of SUV designs, the policy discussions reflect the diversity of potential policy responses. First, to the extent that SUVs generate higher accident costs of a financial nature, this difference will already be reflected to a large extent in insurance rates and will be internalized by SUV owners. Second, there has been litigation in the courts claiming that SUVs do not strike a reasonable balance between risk and other product attributes, thus imposing inordinate risks to passengers in cars. If successful, these lawsuits could establish financial incentives for changes in SUV design. Whether such incentives are desirable, however, is not clear. It presumably would depend on whether there have been inadequacies in consideration of these issues in regulatory contexts by the NHTSA. Third, there may be government regulation of SUV design to address the safety issue. To the extent that vehicle weight is a concern, that risk factor is also present for very large cars. Similarly, the bumper height and stiff frame risk factors are present not just for SUVs but also for light trucks, as they served as the design on which many SUVs are based. If there is a judgment that there is a significant market failure, an efficient policy solution will require a comprehensive analysis of alternative product designs for a broad range of vehicles to strike some balance between these competing effects.

There also may be consumption externalities of a positive nature. It has often been observed that with respect to the use of protective equipment, such as hockey helmets, that individuals are reluctant to use such equipment on their own because doing so would be a sign of weakness in this contact sport. However, if a critical mass of players can be required to wear helmets, then individuals will voluntarily choose to adopt these helmets. This situation reflects an *n*-person Prisoner's Dilemma of the type modeled by Schelling (1978). The observed market equilibrium from individual choices may lead to a less preferred outcome that leads to a lower payoff to all individuals than would a regulated outcome in which people were constrained to undertake behavior that conveyed positive externalities. An example of a policy remedy that will improve

[16] See Gayer (2004) for an analysis of the accident costs associated with SUVs.

social welfare in such situations has been to require the use of hockey helmets and other protective devices for professional athletes.

In some cases, products may generate both positive and negative externalities. A noteworthy example pertains to the external financial costs of cigarettes. Smokers experience morbidity effects and have premature mortality compared to nonsmokers. As a result, there are externalities that increase some social costs, such as medical costs, fires, sick leave, and retirement benefit taxes that are not paid once smokers are dead. Other cost components, notably pensions, social security, and nursing home expenditures, are lower for smokers. On balance, excluding the role of excise taxes, cigarettes convey a net cost savings based on discount rates such as 3 percent or less and impose a net cost increase when the effects are evaluated at interest rates of 5 percent or more. Thus, there are both positive and negative externalities that have different trajectories over time, making the intertemporal weights much more consequential than in simpler externality situations such as with respect to air pollution emissions, which always impose a net external cost.[17] One caveat regarding the self financing aspect of cigarette externalities is that the distribution of the benefits and costs varies depending on the particular externality, so that while on net there may be no net adverse financial externality, the parties responsible for the individual components may not all benefit. For example, even though smoking generates a 32 cent per pack saving in insurance and tax externalities, the present value of the rise in medical insurance costs due to smoking is 58 cents a pack evaluated at a 3 percent interest rate.

One also should be cautious in generalizing from the smoking results to other products. Alcohol use may appear to have similar properties in that excessive alcohol consumption is a risky consumer product. However, as shown by Manning et al. (1989), the substantial costs imposed by drunk drivers are a dominant concern that makes the negative externalities associated with this product loom particularly large.

3.4. Modes of intervention

The existence of market failure need not imply that all possible interventions will be successful in improving market outcomes. The inefficiency that is present may be small, or it also could be the case that no intervention would be beneficial. However, if there is a desire for intervention, there are typically a variety of different forms that this intervention could take. Many of these issues pertain to the choice between regulation and litigation, which has been examined by Shavell (1987).[18] However, even within a particular institutional structure, such as regulation, there may be different types of actions the regulator might adopt.

Consider a consumer's decision to purchase a dangerous product for which the actual risk is p^* and the perceived risk is p, where $p < p^*$. Let the cost of the product be c and

[17] These results are reflective of the findings in both Manning et al. (1989) and Viscusi (2002).

[18] For a comprehensive assessment of the role of the liability system in a wide variety of contexts, see Shavell (2004).

the consumer's income be y. The consumer receives utility from consuming the product of $u(y-c)$ if the product doesn't cause injury, and $v(y-c)$ if the product does cause injury. Alternatively, the consumer could choose not to buy the risky product and reap utility $x(y)$. Suppose too that there is a social cost g generated by adverse outcomes associated with the product's use. Thus, g occurs with a probability p^*.

The consumer will purchase the risky product if

$$(1-p)u(y-c)+pv(y-c)>x(y). \tag{7}$$

This purchase will be socially desirable if

$$(1-p^*)u(y-c)+p^*v(y-c)-pg>x(y). \tag{8}$$

For this consumer choice problem there are consequently two potential sources of market failure creating a disparity between private decisions and social valuations.

To remedy the disparity between p and p^*, the government could simply supply information. If the market failure is with respect to imperfect information, then information disclosure is a natural remedy, as will be discussed below in section 4. The impetus for this informational intervention could come voluntarily through market forces, but if these are inadequate, then either government regulation can require the information disclosure, as with nutrition labeling requirements, or there might be tort liability for firms that fail to disclose the pertinent risk information, as in failure to warn cases.

A second form of intervention that can be useful in aligning private and social incentives consists of tax or penalty schemes. Thus, a penalty value of h can align private and social values for the consumer decision by setting

$$\begin{aligned}&(1-p)u(y-c-h)+pv(y-c-h)\\&\quad=(1-p^*)u(y-c)+p^*v(y-c)-p^*g.\end{aligned} \tag{9}$$

Whereas providing accurate information can align p and p^* but not address the externality g, the penalty h can discourage consumer purchases in much the same way as would greater risk beliefs as well as imposing penalties, so that the external costs are incorporated in the consumer's decision. Thus, for our consumer choice problem, the regulation could be structured to provide the risk levels that people would choose if they were fully informed of the product risks and took full account of external costs they imposed with their choices. The various kinds of "sin" taxes, such as excise taxes on gasoline, alcohol, and cigarettes, are perhaps the most prominent of these types of financial incentive devices.

A third general mode of intervention is the use of government regulation. The form of such regulations can be quite diverse, as they might be technology-forcing regulations as in requirements with respect to the design of buses imposed by the U.S. Department of Transportation, or they might be performance-oriented requirements, as is the case with respect to permissible exposure limits for airborne carcinogens in the workplace or permissible levels of grain dust in grain elevators. Government regulation also may involve the use of monetary incentives and market-based systems, as discussed in the chapter by Revesz and Stavins.

The tort liability system also plays an instrumental role in many instances. As with most forms of government regulation, the financial penalties generated by tort liability are only triggered once the product risk falls below some critical level that ideally reflects some appropriate balancing of cost and risk. Liability costs often are substantial and will provide the same kinds of incentives for safety as would the financial incentives generated by regulation. For example, workplace fatality risks are reduced by about one-third due to the financial incentives provided by the U.S. workers' compensation system.[19] Similarly, product liability costs affect incentives as well, as evidenced in the increased cost of vaccines in response to tort liability for these products.[20] These financial incentives in turn affect incentives for innovation, as small and moderate levels of liability costs increase firms' levels of research and development, product innovation, and patents for new products, which is in line with the conventional economic theory of deterrence.[21] At extremely high levels of liability, however, these incentives may become counterproductive. Thus, firms may simply leave the market or stop producing a particular product if liability costs become excessive.[22] The National Academy of Sciences (1990) attributed the exiting of firms from research on contraceptive devices to the chilling effect of tort liability.

A potential deficiency of ex post remedies such as tort liability is that the injuring party may be judgment-proof, especially for very large losses. For individuals, the result has been that states require mandatory automobile insurance so that parties who are victims of auto accidents can potentially recover damages from an injurer who otherwise would be insolvent. For corporations and other institutional entities, the judgment proof problem arises for catastrophic risks, making it desirable, for example, to have mandatory insurance. Whether safety incentives are created by insurance will depend on the extent of experience rating.

Liability has an additional disadvantage in terms of its ability to replicate the outcomes that would occur with perfect markets. Tort compensation for personal injuries generally provides for economic damages and coverage for noneconomic loss that is below the levels of the value of statistical life pertinent for establishing efficient incentives for accident prevention. Ex ante regulatory interventions can generate such incentives without the problem of providing excessive levels of insurance.

[19] See Moore and Viscusi (1990) for an analysis of wage offsets from workers' compensation.

[20] See Manning (1994) for an assessment of the higher prices resulting from vaccine litigation.

[21] See Viscusi and Moore (1993), who analyze the effect of liability costs on patents and new product introductions.

[22] See Huber (1988).

4. Hazard warnings and risk communication policies

4.1. Principles for hazard warnings

The principal alternative to direct government regulation is the use of informational policies. Whereas government regulation often imposes restrictive limits on individual behavior, hazard warnings work through market processes and seek to address the informational inadequacy directly. Thus, people will then be able to choose the risky products or jobs that are consistent with their own preferences.

A separate rationale for hazard warnings policies is that often it is difficult to have command and control regulations for decentralized behavior. For example, consumers' use of dangerous pesticides is generally not monitored by the government but is undertaken without any direct supervision of the pesticide usage. In such contexts, hazard warnings serve to provide the information that people need in order to take appropriate precautions with respect to usage of the product. If the risk of misuse is especially great, then for these especially dangerous pesticide products the government frequently requires that the user be a certified pesticide applicator. Thus, rather than simply relying on warnings, the government requires additional training programs, but these programs are also informational in character. However, they are more intensive than on-product warning labels.

From the standpoint of efficient market operation, hazard warnings should provide new information in a convincing manner. The objective of warnings is to enable people to have accurate risk beliefs rather than to simply deter behavior. In that regard, information that simply reminds consumers of what they already know generally does not alter behavior and does not serve a constructive function. A prime example of such a reminder warning was the Buckle Up for Safety seatbelt campaign, which had a very disappointing effect on seatbelt use and traffic safety.[23]

To see how these principles stem from an optimal Bayesian learning model, consider the following formulation. In the following example, risk information can be constructive. For simplicity the formulation will be in terms of a single choice that poses the risk of an accident. Suppose as before that individuals have prior risk beliefs p of some adverse outcome, with associated precision γ, where $p < p^*$, which is the true level of the risk. Suppose that the hazard warning conveys risk level q with informational content ξ. Then after people have received the warning they will have posterior risk beliefs p' given by

$$p' = \frac{\gamma p + \xi q}{\gamma + \xi}, \tag{10}$$

assuming rational Bayesian updating. If warnings are not credible and have little information content, or a low value of ξ, then they will not alter risk beliefs and will play

[23] For an evaluation of this effort, see Adler and Pittle (1984). These authors document other failures of informational programs as well.

no constructive role. For very informative warnings, with high values of ξ relative to γ, a value of q equal to p^* will lead people to have accurate risk beliefs. Reminder warnings, such as those with p equal to q, will not alter mean risk beliefs. They will not discourage risk-taking behavior since their only effect is to increase the informational content to γ to ξ, thus leading people to hold these risk beliefs more tightly. To the extent that ambiguity aversion leads people to avoid risk taking when risks are ambiguous, this greater precision will lead people to incur more risks rather than fewer risks if mean risk beliefs are unaffected.

In view of the proliferation of hazard warnings, it is worth noting too the potential shortcomings of excessive hazard warning information. The guiding principle for all informational interventions is that they are subject to individuals' cognitive limitations. People have limited information processing capabilities. As Bettman, Payne, and Staelin (1987) observe, limitations on people's working memory make it important for warnings to be clear, well organized, and readily processed.

Conceptualization of these types of concerns have led to empirical studies of how these limitations affect the practice of effective warnings design. First, researchers have identified that there may be significant problems of information overload.[24] If people are inundated with too much information, then the public may not be able to distinguish the most important risk messages that they receive. If, for example, everything in the supermarket is labeled dangerous, then it will not be feasible for consumers to make relative risk judgments. The second problem is that of label clutter. Although economists frequently claim that more information is more desirable than less information, this principle does not necessarily extend to the practical world of warnings. More information may make it difficult for people with limited information processing capabilities to acquire the information and form reasonable probabilistic judgments.

4.2. *Responses to multiple information sources*

Often people are confronted not with a single information source but multiple information sources regarding a particular risk. In such situations involving a risk debate, how might people form their risk judgments? Using the same notation as before, let q_i be the risk implied by information source i and ξ_i be the associated informational content of information source i, $i = 1, 2$. The credibility of each source and the risk level implied by their warning will vary with the type of risk, the identity of the source, and the basis for the risk judgment that is presented. Then in this instance, posterior risk beliefs will be given by

$$p' = \frac{\gamma p + \xi_1 q_1 + \xi_2 q_2}{\gamma + \xi_1 + \xi_2}. \tag{11}$$

The effect of risk debates on the public's risk beliefs may be complex, and will depend on how people interpret q_i and ξ_i in the context of a risk debate. Two effects have been

[24] Many of these issues are discussed in Viscusi and Magat (1987) and tested in Magat and Viscusi (1992).

established empirically. First, the existence of a risk debate may create a situation of risk ambiguity. Substantial expert disagreement regarding the risk can lead people to assess the risk as being greater than the midpoint value of the two risk experts, assuming that they have equal credibility. Thus, people believe that the ambiguous risk implied by these experts is equivalent to some precisely understood risk level s that is greater than the weighted average of the two risk experts' views. Second, situations in which there is a risk debate involving different entities are especially likely to generate a situation in which the equivalent precisely understood risk is much greater than the weighted average of the risk expert assessments, where these weights are ξ_1 and ξ_2.[25] This result holds even if the divergent risk experts' identities are reversed. Thus, if the government assesses the risk as being high and the polluting industry claims the risk is low, people tend to believe the high-end estimate. Similarly, if the industry expert assesses a high risk level and the government assesses a low risk level, people likewise assess the risk as being closer to the high-end estimate.

4.3. A brief history of warnings

Although we now take hazard warnings policies for granted, in terms of the overall history of risk regulation hazard warnings are a relatively new development. Until early in the twentieth century, there was no legislation whatsoever pertaining to hazard warnings and also no regulatory requirements imposed by regulations. The first such legal requirement pertaining to warnings was the enactment by Congress of the Federal Caustic Poison Act in 1927. For the first time, manufacturers would be required to place hazard warnings on the twelve most dangerous chemicals. This group includes chemicals such as hydrochloric acid and sulfuric acid. After this act, these chemicals would now have to be labeled "Poison."

Perhaps the most ubiquitous warnings today are those for various kinds of pharmaceutical products. Even these warnings are a relatively recent event, as the first food and drug warnings began with the passage of the Federal Food, Drug, and Cosmetic Act in 1938. Whereas warnings today for prescription drugs often focus on long-term effects and adverse drug interactions, the emphasis of these warnings requirements was on acute toxicity. The main concern was with misbranding of products and adulteration of products. The kinds of safety and efficacy concerns that are paramount in today's warnings were simply not on the warnings agenda at that point.

About a decade later Congress extended these warnings requirements to include other products with the passage of the Federal Insecticide, Fungicide, and Rodenticide Act in 1947. For the first time, insecticides and herbicides would be required to have hazard warning labeling.

In the 1960s there was a more rapid extension of these warnings policies. In 1960, Congress passed the Federal Hazardous Substance Labeling Act. This act introduced the

[25] These influences are examined in Viscusi (1997, 1998).

hazard warnings vocabulary that we now take for granted. In particular, it specified the appropriate use of the human hazard signal words, “Danger,” “Warning,” and “Caution.” In addition, this legislation also specified the first requirements that must be met for the warnings associated with substances that are flammable or radioactive.

Also in the 1960s Congress passed warnings legislation for cigarettes in 1965. Beginning in 1966, cigarettes would be required to bear an on-product warning. These warnings requirements were revised in 1969 and again in 1984, at which time Congress specified a series of rotating warnings that remain in place today.

Researchers in the warnings field frequently refer to the 1980s as the “right-to-know” decade. It was in that decade that the Occupational Safety and Health Administration initially imposed hazard communication requirements for dangerous chemicals used in the workplace. As part of this initiative, OSHA also required that material safety data sheets be provided for dangerous chemicals so that it would be possible to treat exposed workers appropriately. Warnings efforts of other kinds also proliferated, as the U.S. Environmental Protection Agency (EPA) began to provide information regarding toxic exposures, and there was the beginning of hazard warnings for a wide range of products, such as lawn mowers. In addition, Congress enacted a series of on-product warnings for alcoholic beverages that were patterned to a large extent on the already-existing cigarette warnings. The main difference was that alcoholic beverage warnings contained multiple warnings on each container rather than a rotating series of single warnings.

The court system also came into play with respect to hazard warnings over the past 40 years.[26] Increasingly firms became liable for their failure to warn consumers adequately of the risks associated with the product. Thus, the incentives for providing warnings came not only through the market and regulatory requirements, but also from legal requirements as well.

5. Risk assessment[27]

5.1. Objectives

If we assume that the policy focus is on expected costs and benefits, then the probabilities used in this calculation should be the mean risk values, as those are used in taking the expectation. Thus, the task is to calculate the expected number of adverse outcomes that will result from a policy. As the discussion below will indicate, many government agencies do not adhere to this approach but instead use upper bound estimates of the

[26] For a discussion of the development of hazard warnings legal requirements, see the American Law Institute (1991).

[27] For further exploration of the risk assessment issues beyond what appears in the following section, see the European Commission (1996), National Research Council (1994, 1996), and the U.S. Environmental Protection Agency (1988, 1996).

component parameters used in calculating the risk, thus yielding a risk assessment that may be in the upper tail of the distribution of what the actual risk might actually be.

The 1983 guidelines for risk assessment issued by the National Research Council (NRC), *Risk Assessment in the Federal Government: Managing the Process*, established standards for risk assessment but did not resolve all outstanding issues. The report claimed, as seems reasonable, that "government regulation rests on the best available scientific knowledge" (NRC, 1983, p. 1). What the "best available scientific knowledge" is that should guide policy is a matter of debate. A principal benefit of the report was shifting the focus away from use of short-term, acute animal studies, which agencies formerly had used in assessing the levels of chemical exposures for which there were no-observed-effects (NOELs). Thus, the emphasis had not been on ascertaining the risk probabilities or even the dose-response relationship, but rather the exposure level at which the risk was no longer zero. More specifically, the focus was on exposure levels for which there is a "reasonable certainty of no harm." The actual NOEL value that agencies set is more stringent. After determining the scientifically valid NOEL amount, agencies then divide this zero observed risk exposure level by 100. Agencies assert that such an adjustment is warranted based on the assumption, for which there is no evidence, that humans are 10 times more sensitive to chemicals than animals, and that there is sufficient risk heterogeneity among people so that the personal risks may differ by a factor of 10 for any given chemical exposure. Moreover, these adjustment factors are assumed to be multiplicative. The EPA has used the NOEL value as its no-observed-adverse-effect level (NOAEL) criterion for evaluating regulatory policies.

In some instances, the policy approach to risk assessment is quite simple. The Food and Drug Administration (FDA) banned all artificial carcinogenic food additives (under the Delancy Clause) irrespective of the magnitude of the risk. Thus, the fact that the additive was artificial was the key concern, not whether the additive posed any real risks or whether the level of carcinogenicity exceeded that of naturally occurring carcinogens in the product.

At least in the case of some agencies' legislative mandates, agencies must show that the magnitude of the risk is consequential before regulation is warranted. In the U.S. Supreme Court decision dealing with the OSHA benzene regulation, the court ruled that the agency must show that the risk is "significant" before regulation is warranted.[28] How large a risk must be to be termed "significant" was not specified by the Court, though the court did observe that a one in a billion cancer risk from a glass of water would not be significant.

In the mid-1980s there was further clarification of the guidelines that risk assessments should meet by the Office of Science and Technology Policy (OSTP, 1985). The risk assessment guidelines issued by EPA (1987) and other agencies were in response to these recommendations. For concreteness, the discussion below will focus on the risk assessment practices for carcinogens, as these risks play the dominant

[28] See *Industrial Union Department, AFL-CIO v. American Petroleum Institute*, 448 U.S. 607 (1980).

role in the estimated regulatory benefits. The agency has no guidelines for assessing the magnitude of morbidity risks so that cancer risks dominate the policy assessments.

5.2. EPA risk assessment practices

EPA relies on both epidemiological studies of human populations and animal bioassay studies in assessing the risk. In most cases the agency relies on animal studies due to the paucity of sound epidemiological studies. Although EPA (1987, p. 1.6) espouses a commitment to statistical significance in this test data, such claims lack content unless one specifies the required significance level. The particular level of significance used for the test is "a matter of overall scientific judgment," according to EPA. The level selected by the agency need not, for example, be based on the usual statistical standard of a 95 percent confidence interval, two-tailed test.

The manner in which EPA combines the results from multiple studies does not place an equal weight on the various studies or weight them based on sample size. Instead, studies indicating the presence of a risk receive greater weight than do negative studies that do not indicate a risk. The result is that, for epidemiological studies, the weight of evidence "increases with the number of adequate studies that show comparable results or populations exposed to the same agent under different conditions" (EPA, 1987, p. 1.6).

EPA then categorizes carcinogens based on different groupings. Group A and Group B carcinogens are either designated human carcinogens or possible human carcinogens. Chemicals in these categories are candidates for risk assessment, as are Group C carcinogens which are possible human carcinogens. Chemicals in Groups D and E are not designated as human carcinogens and, as a result, are not candidates for risk assessment. For chemicals in Groups A, B, and C, EPA may then assess a dose-response relationship. These assessments typically involve extrapolations from animal studies and as a consequence assume that animal carcinogens are human carcinogens and that the evidence for animals can be extrapolated in a meaningful way to humans. The animal studies often focus on the maximum dose of a chemical that can be tolerated. Although the NRC believes that a threshold model is more plausible, EPA uses a linearized model which, in EPA's view, establishes "a plausible upper limit to the risk that is consistent with some proposed mechanism of carcinogenesis" (EPA, 1987, p. 1.9).

People can be exposed to risks in a variety of ways. EPA distinguishes three principal mechanisms of exposure: ingestion, inhalation, and dermal penetration. In some instances there could also be radiation risks. For each of these exposure pathways there is an assessment of the associated risks based on the chemical concentration, the duration of the exposure, and the frequency of exposure. The average daily dose (DOSE) of a chemical is given by

$$\begin{aligned} \mathit{DOSE} = {} & (\mathit{Concentration} \times \mathit{Ingestion\ Rate} \times \mathit{Duration}) \\ & /(\mathit{Weight} \times \mathit{Average\ Time}), \end{aligned} \tag{12}$$

where *Concentration* is the chemical concentration, *Ingestion Rate* is the ingestion rate of the chemical, *Duration* is the exposure duration, *Weight* is the body weight, and *Average Time* is the averaging time, which is a normalization factor. While EPA focused on the maximally exposed individual until 1992, since then it uses what it designates as the high end exposure estimate, which is a "plausible estimate of the individual exposure for those persons at the upper end of an exposure distribution" (EPA, 1992, p. 45).

The EPA risk assessments embody a variety of forms of conservatism. The analysis uses high-end assumptions for each of the parameters in the numerator of the *DOSE* calculation. The practical result is that EPA assessments of the magnitude of the risk are often at or above the 99th percentile of what the actual value of the risk may be.[29] For example, if each of the parameters in the numerator of the *DOSE* calculation above are at the 95th percentile of these distributions and if the distributions are independent, then the chance that the risk is as great as the value EPA calculates is $(.05)(.05)(.05) = 1.25 \times 10^{-4}$, which is far smaller than the 5 percent chance that any of the risk parameters in the calculation are as great as EPA estimates.

Numerous other forms of conservatism bias are incorporated into the analysis. For the animal studies used in calculating the dose-response rates, EPA relies on "long-term animal studies showing the greatest sensitivity" (EPA, 1987, p. 1.8), rather than average animal studies. The risk assessments also assume a level of exposure to the chemical or hazard that is typically a high-end estimate, such as the 90th percentile of the distribution. The default exposure amounts also tend to be at the 90th percentile, such as less than 30 years of continuous exposure to soil ingestion at a hazardous waste site. In some instances, the analysis incorporates 95th percentile values of these parameters.

Many empirical estimates have been developed of the effect of this cascading conservatism.[30] Some early studies, such as those by the Chemical Manufacturers Association (1991) and the Michigan Manufacturers Association (1993) pegged the estimated risk level as being greater than the 95th percentile of the risk distribution. Other Monte Carlo studies, such as those by Cullen (1994) and Finley and Paustenbach (1994) estimated that the EPA-estimated risk levels were greater than the 99th percentile of the risk distribution and, in the case of groundwater contamination, beyond the 99.99th percentile of the risk distribution. Similar results beyond the 99.99th percentile were found in a study of assessments for 141 Superfund sites by Viscusi, Hamilton, and Dockins (1997). That study addressed only the conservatism bias in the site-specific chemical values. Taking into account the upward biases in the dose-response relationship would push the extent of the conservatism bias even higher. The degree of bias varies depending on the policy

[29] For documentation of this effect, see Hamilton and Viscusi (1999) and Viscusi, Hamilton, and Dockins (1997).

[30] Burmaster and Harris (1993) were among the first to explore cascading conservatism biases.

context as well as on the chemical involved since increases in the number of species tests raise the level of bias (see Nichols and Zeckhauser, 1986).

In terms of policy effects, the conservatism practices in many respects embody the same types of influences captured by the Ellsberg Paradox. The government will place greater emphasis on dimly understood risks, compared to mean risk levels. A variety of critics, such as Nichols and Zeckhauser (1986), deplore this bias as distorting policies away from an emphasis that would produce the greatest net expected benefits. Defenders of conservatism, such as Finkel (1989), suggest that such conservatism biases offset biases in the opposite direction, such as the neglect of synergistic effects in risk assessments. However, no empirical evidence has appeared in the literature linking the extent of conservatism bias with the magnitude of any synergistic effects.

The assumption for my discussion has been that the government should be risk-neutral in assessing the expected benefits of regulations. This assumption does not imply that risk attitudes of those being protected are assumed to reflect risk neutrality or to have other properties that government entities might have. The value attached to any given benefit outcome is the individual's willingness to pay for the risk reduction, including whatever values or risk aversion that these preferences may entail. What if, however, a policy offers a 0.5 chance of preventing 10 cases of cancer and a 0.5 chance of preventing 100 cases of cancer? Should that policy be viewed as being equivalent to a policy that prevents 55 cases of cancer? These benefit assessments could differ if, for example, clustered deaths are more highly valued than an equivalent total of individual deaths, but recognizing the role of clustered deaths is a quite different matter analytically than using upper bound estimates of each parameter when doing a risk assessment.

The policy emphasis of these conservatism practices is to some extent similar in spirit to the precautionary principle, which has played a prominent role in European regulatory policies. Interpretation of what is meant by the precautionary principle is not always well defined, and this principle varies across countries, as analyzed by Gollier and Treich (2003) and by Löfstedt (2004). A common version of the precautionary principle is that industry must show that a chemical is safe before it can be used. Thus, the approach is analogous to the tests the FDA imposes on new drugs in the U.S., which is that they must meet tests of safety before the drug can be marketed.[31] From a statistical standpoint it is not feasible to prove that a chemical is safe, but based on tests of the chemical it is possible for one not to reject the hypothesis that there is zero risk. As a practical matter, the precautionary principle permits fewer risk tradeoffs than does a policy based on a comparison of overall benefits and costs. Many European countries are now shifting away from a precautionary principle approach to a broader regulatory impact assessment that is similar in spirit to a benefit-cost test.[32]

[31] Drugs must also meet tests of efficacy as well.

[32] Löfstedt (2004) reviews the evolution of these policies in detail.

6. Valuing the benefits of regulation

6.1. Risk valuation

From the normative standpoint of assessing economic benefits for government policy, the task is to assess society's total willingness to pay for the risk reduction. This formulation is a natural approach for economists, since it follows the general principles for benefit assessment for all government policy.[33] Thus, to monetize the benefits from reduced fatalities one could use the VSL estimates based on fatality risk-money tradeoffs. This approach was a departure from the approach taken in tort liability cases for which the measure in cases of wrongful death is the present value of the decedent's lost earnings, net of consumption. Other factors such as noneconomic losses may enter court awards as well, but for the most part compensation for injuries is structured to serve an insurance function—to compensate the survivors after the fatality. In the case of government regulation, there is no compensation provided, but instead the focus is on striking the appropriate cost-risk tradeoff for purposes of risk reduction, which is a quite different matter from compensation. Thus, the difference between the courts and regulatory agencies in the approaches to valuing life stems more from the different objectives of these social institutions rather than any superiority of one measure versus another.

Although the risk-money tradeoffs reflected in willingness to pay values for risk reduction are the dominant approach in the literature, government agencies often depart from this approach by not examining total benefit values to society. Instead, the emphasis may be narrower. Rather than assessing society's willingness topay for risk reductions affecting the total population, agencies usually use an individual risk approach. Thus, the key concern is whether any individual is exposed or hypothetically might be exposed to a risk level exceeding some threshold, such as a lifetime cancer risk of 1/100,000. Moreover, this hypothetical exposure could come from a quite different hypothetical use of the land. Justice Breyer (1993), for example, expressed his incredulity that EPA wished to clean up a hazardous waste site so that it would be safe enough for children to eat the dirt over 200 days per year instead of only just over 60 days per year. He was puzzled because the site was currently a swamp and there were no dirt-eating children in sight. The vantage point adopted for the discussion here is an assessment of society's willingness to pay for risk reduction for the total population.

The appropriate methodology for conceptualizing risk reduction benefits can be illustrated using risks to life as the example.[34] If a person is willing to pay $700 to eliminate a mortality risk of 1/10,000, then the value of statistical life is $700/(1/10,000) =

[33] Early articulations of how Schelling's (1968) principles for private risk valuation could be extended to the public domain were written by Mishan (1971) and Zeckhauser (1975).

[34] For detailed reviews of this literature and an introduction to this topic, see Jones-Lee (1976, 1989), Smith (1979a), and Viscusi (1992, 1998). International evidence appears in Kniesner and Leeth (1991). Viscusi and Aldy (2003) provide the most recent review.

$7 million. This value can then be used in benefit assessments and, in much the same way, these values can be constructed for other risk reduction outcomes assuming that the benefit transfer is warranted. Thus, the risk preferences, baseline risk levels, and extent of the risk reduction should be similar, or the VSL levels should be adjusted appropriately. Following the general principles of the public finance literature, one can likewise develop willingness-to-pay values for benefit components other than mortality risks. Such values are more appropriate than, for example, relying on medical expenses alone to value risks of injury or using travel costs to value recreational benefits of environmental amenities.

The willingness-to-pay values are for saving statistical lives, not identified lives for which the risk of death is reduced from 1.0 to zero. Jenni and Loewenstein (1997) distinguish four aspects of contexts in which the victims are identifiable. First, the deaths involve the certainty of death rather than variations in small probabilities of death. Second, identifiable deaths are often vivid and highly publicized specific individuals, generating substantial public concern. Third, the fraction of the reference group that will be saved is much greater than in a typical risk regulation program, possibly as high as 100 percent for an individual at risk. Fourth, in the case of identifiable victims, the decision is made after the lottery has been run rather than before, making it an *ex post* decision rather than an *ex ante* decision.

How we conceptualize benefit values also depends on the scope of the willingness-to-pay measure. The appropriate benefit values for reducing risk are structured based on the person's willingness to pay for private benefits, but altruistic concerns may enter as well. People may be willing to pay for risk reductions for others. How and whether such valuations should enter depends on the source of the altruism. The model by Jones-Lee (1991) indicates that safety-based altruism that is independent of utility levels should be recognized as a legitimate benefit component.[35]

In the case of morbidity effects, much of the emphasis in the literature has been on estimating the willingness to pay to reduce the risks of illness.[36] The traditional human capital approach sets the benefit value equal to the opportunity costs of being sick plus the costs of mitigating behavior as reflected, for example, in medical expenses. Expenditures on behavior directed at averting risks and the disutility of being sick must also be included to have a comprehensive benefit measure. These various measures of a willingness-to-pay approach are generally substitutes for human capital benefit measures, not additive components. Other measures of morbidity risk benefits such as those implied by wage-risk tradeoffs in the labor market and survey valuations may produce comprehensive morbidity benefit values.[37]

[35] Interestingly, a test of the Jones-Lee (1991) model by Johannesson, Johansson, and O'Conor (1996) failed to show that respondents generally would pay a higher value for public safety than private safety improvements.

[36] Harrington and Portney (1987) and Freeman (1993) developed much of the underlying analysis for the morbidity benefit estimation approach.

[37] It is interesting to note that, while one might think that only men are exposed to job risks, women face substantial injury risks as well and also exhibit wage-risk tradeoffs similar to men, as shown in Hersch (1998).

Most quantified benefits of risk and environmental regulation are from studies of mortality risk reduction, which in turn are usually based on labor market studies of the value of statistical life. These have, of course, been efforts to impute the VSL for environmental contexts by linking housing process to pollution levels, as in Portney (1981). However, this literature is much less extensive and often involves stronger assumptions in the construction of the risk variable than do labor market studies.[38] Housing has been a chief market context for analyzing price-fatality risk tradeoffs, but there have been other product market analyses as well, especially automobiles.[39]

Suppose that $w(p)$ represents the market offer curve of the highest wage rate available to workers at any given mortality risk p. This curve is the outer envelope of firms' isoprofit curves. Firms will offer a lower wage as the safety level increases since safety improvements are costly, and the wage rate associated with greater safety levels must be less to keep the firm on the same isoprofit curve.

Workers choose the point along this offer curve that provides them with the greatest expected utility. Here we adopt a simplified version of the formulation above as utility only depends on the worker's wages and not also on assets and health expenditures. Let the state-dependent utility function $u(w)$ be the worker's utility when healthy and $v(w)$ be the utility function when the worker is injured or the bequest function in the case of fatality risks. As in the model in section 2, these utility functions have the usual properties $u(w) > v(w)$; $u'(w)$, $v'(w) > 0$; and $u''(w) \leqslant 0$, $v''(w) \leqslant 0$. Note that financial risk aversion is not required for workers to find risky jobs unattractive. For any given wage rate, increasing the risk p is always undesirable since good health is preferred to ill health.

The optimizing worker will pick the point off the market opportunities frontier that provides the greatest expected utility. At this point the worker's highest valued constant expected utility locus will be tangent to the market offer curve, with the wage-risk tradeoff at this point of tangency being the worker's value of statistical life. The wage-risk tradeoff at this point is given by

$$\frac{\partial w}{\partial p} = \frac{u(w) - v(w)}{(1-p)u'(w) + pv'(w)}, \tag{13}$$

which is a special case of the more general result in equations (3) and (4). Thus, the implicit value of a statistical life equals the difference in utility levels between the healthy state and the ill state divided by the expected marginal utility of income. One can develop a similar analysis for other market-based tradeoffs, such as those involving

[38] Viscusi and Aldy (2003) provide a review of many published housing price VSL studies, which use either air pollution risks or hazardous waste risks in the analysis. An early study of VSL amounts implied by landfill risks was Smith and Desvousges (1986). Responses to natural disasters have also been the subject of scrutiny, as in Brookshire et al. (1985).

[39] The first of many such studies showing that consumers pay more for safer cars was by Atkinson and Halvorsen (1990).

product choices or housing choices, where these VSL amounts will be the same as those reflected in the labor market.

While these tradeoffs will be pertinent for small changes in the risks of death, they will overstate the amounts that people are willing to pay for large improvements in safety and will understate how much people must be compensated for large increases in risk. The underlying influence responsible for this result is that there are wealth effects with respect to people's preferences with respect to changes in safety levels. Increases in wealth will make people more reluctant to bear risk given the assumptions above.

Empirical implementation of this approach has consisted of hedonic wage and hedonic price studies that trace out the locus of market equilibria consisting of a series of observed wage–risk combinations observed in the market. Thus, the hedonic wage equation does not trace out either the supply of workers to risky jobs or market offer curves, but rather the joint influence of these two sides of the market or the locus of observed market equilibria. The review by Smith (1979a), in particular, stresses this distinction, which has played a prominent role in the literature.

The canonical hedonic wage equation is of the form

$$\ln w_i = \alpha + \sum_{j=1}^{n} \beta_j x_{ij} + \gamma p_i + u_i, \tag{14}$$

where p_i is the fatality risk facing worker i, x_{ij} is a series of variables j pertaining to the worker i's personal characteristics and job, where these possibly include measures of workers' compensation and nonfatal injury risk, and u_i is an error term. The median value of the implicit value of life reflected in studies throughout the world of wage-risk tradeoffs is $7 million.[40]

The VSL estimates may vary considerably based on the risk preferences of the workers. The early estimates by Thaler and Rosen (1976) focused on workers facing risks about an order of magnitude greater than the national average, leading to lower estimated VSL levels than in other studies. This low VSL level is what one would expect if workers who are more willing to bear risk sort themselves into the riskiest jobs.

More recent research has made an explicit attempt to examine the source of the heterogeneity in these values. Some relationships that have been borne out have not been of central interest to policymakers. For example, cigarette smokers are more willing to incur risks on the job, and people who use seatbelts regularly are less willing to bear such risks, as shown in Hersch and Viscusi (1990), Hersch and Pickton (1995), and Viscusi and Hersch (2001). Interestingly, these labor market estimates for wage-risk tradeoffs are consistent with the patterns revealed by smokers' decisions in the product market. Ippolito and Ippolito (1984) find that smokers' responses to new risk information implies a VSL level that is below the typical labor market average.

Within the context of hedonic wage and price models there could be two sources of individual differences. First, people may share the same offer curve $w(p)$ but pick

[40] A recent survey of this literature yielding this value is Viscusi and Aldy (2003).

different points along it. For an offer curve of the usual assumed shape, $w_p > 0$ and $w_{pp} < 0$, workers who chose lower levels of risk p will be on steeper portions of the offer curve and will exhibit higher VSL amounts.

An alternative structure that alters the standard hedonic model is that workers may face quite different offer curves and be settling into separate labor market equilibria. If two groups of workers face identical offer curves, then the group incurring the greater risk should receive greater wage compensation for risk. Let there be two groups of workers $i = 1, 2$ who each face risk p_i and earn wage $w_i(p_i)$. Each worker would earn the same amount for a zero-risk job, or $w_i(0) = 0$. Suppose that $p_2 > p_1$ and that the workers face the same offer curve, but that worker 2 receives a lower wage-rate. Then

$$w_2(p_2) - w_2(0) < w_1(p_1) - w_1(0), \tag{15}$$

or, since $w_2(0) = w_1(0)$,

$$w_2(p_2) < w_1(p_1). \tag{16}$$

But under the assumption of identical offer curves,

$$w_2(p_2) = w_1(p_2), \tag{17}$$

which leads to a contradiction of inequality (16).

This generalization of the standard hedonic model facilitates an interpretation of observed empirical results. Smokers incur greater risks than nonsmokers, yet receive lower premiums for risk. Similarly, black workers face larger risks than white workers but are compensated less for these risks controlling for appropriate demographic factors and job characteristics. These findings imply that smokers face different offer curves than do nonsmokers, and black workers face different offer curves than white workers. In the case of cigarette smokers, these workers incur greater risks for less hazard pay than nonsmokers—a result that cannot occur if nonsmokers and smokers faced the same offer curve. Rather, there are separate hedonic market equilibria for different labor market groups. It is not yet clear how or whether this heterogeneity in market valuations for risk should enter the policy debate.

A more controversial current policy issue pertains to the role of individual age. In particular, should reducing risks to the lives of older people be valued less than reducing risks to the young? Older people have a lower quantity of expected remaining lifetime, but that doesn't necessarily imply that valuations peak at birth and decline steadily with age. What matters is their willingness to pay for such risk reductions. The quality of life at risk may also diminish with age, but an offsetting factor is that wealth may increase. Some standardized measures of valuation assume that there is a diminution in the value with age. The Disability-Adjusted Life Year (DALY) approach in the World Health Organization study reported by Murray and Lopez (1996) assumes that the value of life years peaks in one's twenties and declines thereafter. Similarly, straightforward adjustments in the value of life based on remaining life expectancy or discounted expected remaining years of life will also diminish the benefit values associated with reducing risks to older people.

This age-adjustment issue became a policy controversy in 2003 with respect to EPA's benefit assessments with respect to the Clear Skies initiative. Use of a lower benefit value for the elderly became known as the "senior discount" or "seniors on sale, 37% off" and was abandoned in policy calculations after substantial political outcry by senior citizen groups, such as the AARP. However, if one were to use a uniform value of life for all citizens, that would have the appearance of fairness in one sense but might also be viewed as being unfair to the young, since each year of their life would receive a lower value than each year of older citizens' lives. Use of a constant value per year of life has a fairness appeal that may be as great as using a uniform value for a statistical life. Ultimately, the issue is one that is not a question of fairness or symmetry in valuation but how the willingness to pay to reduce fatality risks varies with age.

From an economic theory standpoint, the value of statistical life should eventually decline over the life cycle. Models such as those by Shepard and Zeckhauser (1984) and Johansson (2002) have shown that if capital markets are imperfect, then the value of life will rise and then fall over the life cycle.[41] These models also demonstrate that the value of statistical life at any given age is dependent on current and future consumption levels. Because consumption rises then falls over the life cycle, one should expect a similar temporal pattern in the variation of the value of statistical life with age.

The two approaches to estimating changes in values of statistical life have been to use survey estimates of the willingness to pay for risk reduction and estimates derived from market data to analyze changes in the implicit value of statistical life with age. The willingness-to-pay survey study by Jones-Lee (1989) found a declining value of life with age, whereas the more recent survey study by Krupnick et al. (2002) found a much flatter relationship, though this study considered a narrower age band than did Jones-Lee (1989). Many market-based studies of age-related variations in workers' value of statistical life showed a steadily declining relationship, but these early studies tended to use average risk levels by industry, with no accounting for age variations in jobs.[42] More recent estimates using age-dependent risk measures[43] and analyses that take into account consumption variations over the life cycle[44] each show that the value of statistical life displays the inverted-U shaped pattern predicted by most theoretical analyses, but that the decline in the value of life with age is not steep. Older workers are, for example, much less willing to incur fatality risks than are workers in their early twenties.

To see how age-related difference in the value of life can affect benefit assessments consider the different estimate of the mortality reduction benefits for the U.S. EPA Clear Skies Initiative, which are reported in Table 1. Using a constant VSL of $6.1 million, EPA found that the bulk of the estimated benefits of the policy are for senior citizens, whether the analysis is based on long-term exposure or short-term exposure risk estimates. If one applies the senior discount of 37 percent used in an illustrative analysis

[41] Using a different framework, Rosen (1988) predicts that VSL will decline with age.

[42] These studies are reviewed in Viscusi and Aldy (2003).

[43] See Aldy and Viscusi (2004), which uses fatality risk and injury risk data by age.

[44] Kniesner, Viscusi, and Ziliak (2006) incorporate consumption into the wage equations.

Table 1
Age group effects on Clear Skies initiative benefits

Age group	Reduced annual fatalities in 2010	Benefits of reduced mortality ($ billions undiscounted)	
		Constant value of life	Value with senior adjusted
Base estimates—long-term exposure:			
Adults, 18–64	1,900	11.6	11.6
Adults, 65 and older	6,000	36.6	23.1
Alternative estimate—short-term exposure:			
Children, 0–17	30	0.2	0.2
Adults, 18–64	1,100	6.7	6.7
Adults, 65 and older	3,600	21.9	14.7

Note: The reduced annual fatalities figures are from the U.S. Environmental Protection Agency (2003), Table 16. The 37 percent senior discount is from the U.S. Environmental Protection Agency (2002, p. 35), and the $6.1 million figure per life is from the U.S. Environmental Protection Agency (2003, p. 26).

by EPA, estimated benefits drop by a bit less than that proportion given that most of the mortality reduction benefits are concentrated among those age 65 and older. If, however, one uses VSL amounts that are reflective of the VSL levels based on one set of estimates of age differences revealed through workers' job choices, the estimated benefits are roughly the same as in the case without age adjustments.[45]

6.2. *Contingent valuation and survey methods*

Reliance on market evidence for imputing private valuations is generally preferable to willingness to pay survey questions when reliable market data are available. However, there are many situations in which meaningful market estimates are not feasible. Perhaps the chief example in which such survey methods are used is that for valuation of the losses associated with natural resource damages, such as the harm caused to natural resources. There are also other less dramatic contexts in which survey methods are used to obtain benefit values. Obtaining survey estimates of society's willingness to pay may be the best available benefits approach to assess benefits such as improved visibility at the Grand Canyon, the discomfort of respiratory illnesses, or the value of saving an endangered species. Indeed, to the extent that there is a concern with non-use values or passive use, then use of some kind to survey approach is the only viable technique for deriving empirical estimates of the benefits.

[45] These particular age variations in VSL are estimated by Kniesner, Viscusi, and Ziliak (2006) using a hedonic wage model that incorporates life-cycle consumption patterns into the analyses.

Survey techniques such as contingent valuation methods, stated preference methods, and conjoint analysis serve as the mechanism for obtaining such valuations. These methods have been highly controversial, and in part because of this controversy, considerable attention has been devoted to the development of criteria that reasonable surveys should meet.[46] Thus, while use of stated preference methods of various kinds may not be preferable to market-based estimates when they are available, for many environmental benefits these approaches represent the only technique for developing meaningful benefit estimates. While a comprehensive review of these issues would be quite lengthy, there are several key concerns that must be addressed for such surveys to have validity.

Because these surveys ask people to engage in a hypothetical market transaction, the respondent must understand the good being valued, and the payment mechanism must be realistic. If people treat the survey questions as pertaining to hypothetical interview money and do not recognize that there is a real economic opportunity cost, then valuations will be too great.

A long-term concern in the economics literature has been that people might misrepresent their preferences if a public good is being valued. Such strategic misrepresentation has not proven to be a major problem and can be addressed by, for example, structuring the survey in terms of votes on a referendum and imputing the tradeoff rates statistically using, for example, a random utility model. A more important practical problem is demand effects, as highlighting an environmental amenity in the survey may make respondents willing to have greater stated preferences for that good than if they took into account possible alternative uses of these funds.

In an effort to address the potential biases of hypothetical choice contexts, researchers also have devised a series of rationality tests that surveys should meet, many of which mimic the usual economic criteria for rational consumer choice. Respondents' willingness to pay should increase with the amount of the good provided. Thus, policies that reduce the risk of harm to a greater extent should be more highly valued.

In contrast, some surveys have shown that people are willing to pay the same amount to save 100 birds as to save 10,000 birds.[47] This lack of responsiveness to the amount of environmental good may arise because people suffer from embedding effects.[48] Thus, respondents in each instance may view their response as simply a vote of support for the environment more generally rather than distinguishing the quantity of the good. A basic rationality test that should be met is that across subjects, willingness to pay values should increase as the amount of the good provided increases. However, this increase in valuations need not be linear, so that people shouldn't necessarily value

[46] For an introduction to many of the controversies surrounding contingent valuation, see Diamond and Hausman (1994). Hanemann (1994) and Portney (1994) have a more positive view of developing contingent valuation techniques to provide sound benefit estimates.

[47] See Desvousges et al. (1993) for analysis of the insensitivity of valuations to the extent of environmental improvement.

[48] This term characterizing the limitation of contingent valuation studies was coined by Kahneman and Knetsch (1992).

saving 10,000 birds by 100 times more than they value saving 100 birds. There may be a rapidly diminishing marginal willingness to pay that will generate similar values from an accurate assessment of individual preferences. Another rationality check is that for tests within subjects, the preferences should satisfy the usual economic rationality criteria such as transitivity.

When there are multiple goods involved, there are rationality tests pertaining to the scope of the good provided. If a policy saves birds and fish rather than fish alone, then the willingness to pay for that policy should be greater. This test is simply the analog of the test for single dimension goods in that people should prefer more of the good to less.

In some instances researchers have used a currency other than dollars for the valuation. Thus, for example, respondents might be asked to choose between a reduction of risks of automobile accidents against reduced risks of cancer from environmental pollution.[49] Such tradeoffs can address the problems that might arise if people treat survey interview money as having little value and may also be useful in enabling respondents to compare two commodities that are both non-monetary and involve probabilistic goods. These comparisons of two goods on the same nonmonetary dimension, such as comparing reductions in auto accidents to environmental improvements, can also overcome respondents' reluctance to monetize environmental goods. In addition, if the task requires valuing small risks that would otherwise present risk denominators that were difficult to process, such as a risk of 1/100,000, then these probability values could be made comparable for the two risks being compared so as to reduce the cognitive demands of the valuation task.

These methodologies remain in the developmental stage, though much progress has been made since the contingent valuation debate was launched over a decade ago. No general conclusions are possible regarding the validity of various stated preference approaches. However, it has become increasingly apparent that survey methods often can serve a useful function for assessing many categories of benefits that have no market-based valuations. In addition, even though the validity of these studies must be assessed on a case by case basis, a series of validity tests has been developed that will provide a degree of quality control.

Survey methods are especially useful when market-based estimates are either unavailable or unstable. Thus, these techniques have even proven useful in estimating VSL levels, which is the context in which labor market estimates have been particularly well developed. However, these estimates have not shown the same pattern of stability using data from the U.K. The labor market studies for the U.K. have yielded wildly varying estimates.[50] As a result, policymakers assessing the benefits of transport safety policies rely on survey values such as those elicited in the willingness to pay study of rail travel

[49] Magat, Viscusi, and Huber (1996) use an automobile fatality accident risk metric to value fatal and nonfatal cases of cancer.

[50] These variations are surveyed in Viscusi and Aldy (2003).

and auto travel undertaken by Jones-Lee (1989) and Jones-Lee, Hammerton, and Philips (1985). The VSL estimates derived from these surveys are more in line with U.S. labor market VSL amounts than are the U.K. labor market studies.[51]

7. Benefit-cost analysis

The economic efficiency test for regulations is that at the minimum the benefits exceed the costs and that ideally one should maximize the spread between benefits and costs.[52] Consider a mortality-reducing regulation that only has effects in the current period. Let the cost of the regulation be c, the extent of the average mortality risk reduction be s, the number of people affected be n, and the value of statistical life be v. Ideally, the policy should maximize the spread between benefits and costs, or maximize the difference $snv - c$. Regulators are seldom held to this standard, but instead must meet the requirement that the benefits exceed the costs, or

$$snv > c. \tag{18}$$

This seemingly innocuous test has a strong justification, but is not always compelling. The caveats regarding benefit-cost tests in the public finance literature generally focus on concerns relating to the Hicks–Kaldor compensation criteria, which are that the gainers aren't always compensating the losers, in which case the policy is not particularly attractive to those whose welfare is reduced. The Hicks–Kaldor potential compensation criterion is not as compelling as it would be if such compensation actually were paid.

The departures from the benefit-cost approach in the practice of risk regulation policy are much more diverse. The transportation safety regulations are a notable exception in that they meet a benefit-cost test. The benefit-cost test can be rewritten as

$$s > \frac{c}{nv}, \tag{19}$$

or the risk reduction must exceed the policy cost divided by the population affected multiplied by the value of statistical life. In contrast, risk policies are often governed by the test

$$s > s^*, \tag{20}$$

where s^* is a critical risk probability triggering regulation. In the case of EPA hazardous cleanup efforts, a value of s^* above 1/10,000 makes cleanup mandatory, and a value of s^* such that $1/10{,}000 \geqslant s^* \geqslant 1/100{,}000$ puts cleanup in the discretionary range. Sites posing lifetime risks to at least one current or hypothetical future citizen of magnitudes

[51] They also yield results similar to those in U.S. survey studies, which are reviewed in Viscusi (1992).

[52] Equity concerns often enter policy debates as well, especially with respect to environmental policies' effects by race. See Hamilton (1995) for examination of these issues, as well as Hamilton and Viscusi (1999).

below 1/100,000 are not cleaned up. Note that the population size n does not enter consideration. Thus, policies are often based on criteria that are characterized as the individual risk approach, in which risks to a single individual drive the policy decision. In contrast, the more comprehensive benefit-cost test takes what has been termed the population risk approach in which the number of people whose risks are being reduced is of consequence.

The benefit-cost test can be rewritten in cost-effectiveness terms as

$$\frac{c}{sn} < v, \tag{21}$$

or the cost per life saved is less than some critical value v. This formulation is useful in characterizing the relative performance of regulations, as will be done below. It can also be used in assessing regulatory performance in instances in which one is perhaps reluctant to commit to a particular value of statistical life v.

Justice Breyer (1993) uses the approach of equation (21) in the following manner. Efficient risk reduction will lead people to equate the cost per unit risk across different areas of risk reduction activity. Support for a regulation that would impose an inordinate cost per expected life saved, such as an inefficient asbestos regulation, could only be justified if one were willing to spend an additional $20,000 on a car that was marginally safer. In effect, Breyer contrasts the risk-money tradeoff implied by inefficient regulatory policies with hypothetical but realistic private choices to demonstrate their undesirability.

This approach simply articulates the theoretical underpinnings of the market-based methodologies used to estimate v. These values reflect the price-risk or wage-risk tradeoff estimates based on market decisions, thus establishing v empirically. Breyer instead uses the same comparison implied by equation (21) above, except he does so by indicating a government risk-cost tradeoff that dwarfs people's sense of what v is in their own decisions.

Once the element of multiple time periods is introduced, the benefit and cost components can have a time dimension t. Letting r be the rate of discount, the policy benefit-cost test for efforts that impose an initial cost c at time zero and generate a benefit stream over time is

$$c < \sum_{t=0}^{\infty} \frac{p_t n_t v}{(1+r)^t}, \tag{22}$$

where the value of statistical life v does not vary over time though it could if there are, for example, important income effects.

This condition can be rewritten in a form that provides a convenient index of policy efficacy. In single period contexts the criterion for attractive policies is that the cost per expected life saved be less than some critical value v. With multiple periods this requirement is that the cost per discounted expected life saved be below the VSL, or

$$\frac{c}{\sum_{t=0}^{\infty} \frac{p_t n_t}{(1+r)^t}} < v. \tag{23}$$

In addition to the usual controversies over the appropriate rate of discount, there are also debates over what the discount rate should be, particularly in the case of environmental policies. Some analysts have suggested that there be no discounting at all for environmental benefits. This practice is not only inconsistent with people's revealed rate of time preference in market decisions; it will also lead to anomalies. Any environmental damage that will be inflicted at a level of a dollar per year forever will swamp any finite benefit estimate. Thus, there could never be a policy justification for making extinct a not particularly interesting plant species, because doing so would impose infinite environmental costs, even if there were very large but finite benefits, such as completely eradicating all current causes of premature mortality. In addition, in a world without discounting, if the costs of the policy decrease over time due to technological progress, it will always be desirable to postpone taking action, because the present value of costs will be less if decisions are deferred, and the infinite benefit stream will be unaffected.

The discounting controversy that has achieved greater prominence pertains to the treatment of future generations. Some observers, such as Revesz (1999), have suggested that benefits to future generations not be discounted.[53] What is meant operationally by this proposal is unclear. Suppose that the current generation ends at time T and the future generation begins at time $T + 1$. How is the value of T determined? Is it the life expectancy of the average voter now alive, the life expectancy of the people now being born, or some other value? Once T is determined, does a policy of not discounting benefits to future generations imply that the policy criterion is

$$c < \sum_{t=0}^{T} \frac{p_t n_t v}{(1+r)^t} + \sum_{t=T+1}^{\infty} p_t n_t v, \tag{24}$$

or

$$c < \sum_{t=0}^{T} \frac{p_t n_t v}{(1+r)^t} + \sum_{t=T+1}^{\infty} \frac{p_t n_t v}{(1+r)^{T+1}}? \tag{25}$$

In the former instance, the failure to discount benefits to future generations at all implies that saving any expected lives in future generations receives a greater weight than saving a life for the current generation for every period other than time zero. The second formulation eliminates this problem by ending the discounting based on a discount factor linked to a time $T + 1$. However, there remains the problem that permanent risk reduction effects for future generations will receive an infinite value, thus swamping effects for the current generation. There will also be inequities within future generations that will not be consistent with the policies they will select themselves. Thus, decision makers at time $T + 1$ will discount benefits at time t' ($t' > T + 1$) by $1/(1+r)^{t'-T-1}$.

[53] For a general treatment of the economics of discounting fatality risk benefits, see Cropper and Portney (1990). Horowitz and Carson (1990) present empirical evidence on respondents' willingness to discount statistical lives.

However, if there is no discounting for future generations when decisions are made now, this discount factor with respect to delays within the future generation will be 1.

Discounting also has come under fire with respect to discounting of physical units of environmental amenities. However, what is being discounted is not the physical benefits but rather society's willingness to pay. Thus, it is entirely appropriate to calculate a cost per discounted expected life saved as was done in equation (23) above and compare it to the value of v.

Equivalently, one could convert the cost value to its terminal value at the time the benefits are generated rather than discounting lives. Thus, if z is the expected number of lives a policy costing c saves after a 10 year lag, one could calculate the cost per discounted life saved as $c/(z/(1+r)^{10})$, or instead calculate the terminal value of the cost, $c(1+r)^{10}$ and divide that by the undiscounted value of z. In each case, the cost-risk tradeoff value is $c(1+r)^{10}/z$.

Other standard controversies surrounding benefit-cost tests are that benefit values for goods not traded in the market will be undervalued and costs will be overestimated. The concerns about undervaluing environmental amenities and similar commodities largely predate the development of survey valuation methods, which enable researchers to develop benefit estimates that many view as too high rather than too low. The cost underestimation issue is one of practical policy implementation and raises concerns as to whether policymakers can learn from past biases in cost estimation, should they exist. Companies, for example, have an incentive to overestimate compliance costs, but this self-interest is known to government regulators.

The overall performance of regulations with respect to the cost per life saved is reflected in the statistics in Table 2. This table, which is based on the work of the Office of Information and Regulatory Affairs at the U.S. Office of Management Budget, was reported by Morrall (2003). The table draws on the results in Hahn, Lutter, and Viscusi (2000), which presented estimates of the net costs per life saved, where the costs are net of benefits other than the lives saved. Table 2 lists the opportunity cost per life saved for a large series of major government risk and environmental regulations.

Such compilations of costs per life saved have become controversial in the past, in part because the calculations sometimes have been misunderstood. Consequently, it is worth emphasizing several aspects of the way in which this table is constructed. First, the measure of regulatory efficacy is the opportunity cost per life saved. This figure parallels the more conventional cost per statistical life saved calculation, except that the opportunity cost of funds is brought to the future at a seven percent rate of interest rather than discounting the number of lives saved back to the current period. As was indicated in the discussion above regarding equation (23), this approach is mathematically equivalent to a cost per discounted life saved approach.

The regulations that appear in Table 2 appear in order of the opportunity cost per life saved. The most cost-effective regulations are at the top of the table, and the least cost-effective regulations are at the bottom. In every instance, it should be emphasized that the scale of the regulatory effort also is influential in affecting the total benefits less costs of these efforts. Thus, the ranking is in terms of the rate at which benefits are

Table 2
Opportunity costs per statistical life saved (OCSLS)

Regulation	Year issued	Agency	OCSLS (millions of 2002 $)
Childproof Lighters	1993	CPSC	0.1
Respiratory Protection	1998	OSHA-H	0.1
Logging Operations	1994	OSHA-S	0.1
Electrical Safety	1990	OSHA-S	0.1
Steering Column Protection	1967	NHTSA	0.2
Unvented Space Heaters	1980	CPSC	0.2
Safety Standards for Scaffolds	1996	OSHA-S	0.2
Cabin Fire Protection	1985	FAA	0.3
Trihalomethanes	1979	EPA	0.3
Organ Procurement Regulations	1998	HHS	0.3
AED on Large Planes	2001	FAA	0.3
Mammography Sts	1997	HHS	0.4
Food Labeling Regulations	1993	FDA	0.4
Stability & Control During Braking/Trucks	1995	NHTSA	0.4
Electrical Power Generation	1994	OSHA-S	0.4
Passive Restraints/Belts	1984	NHTSA	0.5
Fuel System Integrity	1975	NHTSA	0.5
Underground Construction	1983	OSHA-S	0.5
Head Impact Protection	1995	NHTSA	0.7
Alcohol & Drug Control	1985	FRA	0.9
Servicing Wheel Rims	1984	OSHA-S	0.9
Reflective Devices for Heavy Trucks	1999	NHTSA	0.9
Seat Cushion Flammability	1984	FAA	1.0
Side Impact & Autos	1990	NHTSA	1.1
Medical Devices	1996	FDA	1.1
Floor Emergency Lighting	1984	FAA	1.2
Crane Suspended Personnel Platform	1984	OSHA-S	1.5
Low-Altitude Windshear	1988	FAA	1.8
Electrical Equipment Sts./Metal Mines	1970	MSHA	1.9
Trenching and Excavation	1989	OSHA-S	2.1
Traffic Alert & Collision Avoidance	1988	FAA	2.1
Children's Sleepwear Flammability	1973	CPSC	2.2
Side Doors	1970	NHTSA	2.2
Concrete & Masonry Construction	1985	OSHA-S	2.4
Confined Spaces	1993	OSHA-S	2.5
Hazard Communication	1983	OSHA-S	3.1
Child Restraints	1999	NHTSA	3.3
Benzene/Fugitive Emissions	1984	EPA	3.7
Rear/Up/Shoulder Belts/Autos	1989	NHTSA	4.4
Asbestos	1972	OSHA-H	5.5
EDB Drinking Water Sts.	1991	EPA	6.0
NO_x SIP Call	1998	EPA	6.0
Benzene/Revised: Coke By Products	1988	EPA	6.4
Radionuclides/Uranium Mines	1984	EPA	6.9

(*Continued on next page*)

Table 2
(*Continued*)

Regulation	Year issued	Agency	OCSLS (millions of 2002 $)
Roadway Worker Protection	1997	FRA	7.1
Grain Dust	1988	OSHA-S	11
Electrical Equipment Sts./Coal Mines	1970	MSHA	13
Methylene Chloride	1997	OSHA-H	13
Arsenic/Glass Paint	1986	EPA	19
Benzene	1987	OSHA-H	22
Arsenic/Copper Smelter	1986	EPA	27
Uranium Mill Tailings/Inactive	1983	EPA	28
Hazardous Wastes Listing for Petroleum Sludge	1990	EPA	29
Acrylonitrile	1978	OSHA-H	31
Benzene/Revised: Transfer Operations	1990	EPA	35
4.4 methylenedianiline	1992	OSHA-H	36
Coke Ovens	1976	OSHA-H	51
Nat. Primary and Secondary Drinking Water Regulations Phase II	1991	EPA	50
Uranium Mill Tailings/Active	1983	EPA	53
Asbestos	1986	OSHA-H	66
Asbestos/Construction	1994	OSHA	71
Arsenic	1978	OSHA-H	77
Asbestos Ban	1989	EPA	78
Ethylene Oxide	1984	OSHA-H	80
Lockout/Tagout	1989	OSHA-S	98
Hazardous Waste Management/Wood Products	1990	EPA	140
DES (Cattlefeed)	1979	FDA	170
Benzene/Revised: Waste Operations	1990	EPA	180
Sewage Sludge Disposal	1993	EPA	530
Land Disposal Restrictions	1990	EPA	530
Hazardous Waste: Solids Dioxin	1986	EPA	560
Prohibit Land Disposal	1988	EPA	1,100
Land Disposal Restrictions/Phase II	1994	EPA	2,600
Drinking Water: Phase II	1992	EPA	19,000
Formaldehyde	1987	OSHA-H	78,000
Solid Waste Disposal Facility Criteria	1991	EPA	100,000

Source: Morrall (2003). S = safety regulation, H = health regulation.

generated relative to the cost, but are not necessarily in order of the overall net benefits to society.

In an earlier version of this table prepared by Morrall (1986), the table listed final, proposed, and rejected rules, and indicated on the table the status of each rule. The tabulation in Table 2 of 76 regulations focuses only on final regulations and does not include rejected or proposed regulations.

What is perhaps most striking about this table is the difference across agencies in the relative cost-effectiveness of their efforts. For agencies within the U.S. Department of Transportation—the Federal Aviation Administration (FAA) and the National Highway Traffic Safety Administration (NHTSA)—all regulatory proposals meet a benefit-cost test in that none of them imposes an opportunity cost per life saved in excess of $4.4 million. In contrast, many of the regulations at the bottom of the table impose inordinately high costs per life saved. This distribution is largely attributable to the difference in the agencies' legislative mandates, as the requirements for EPA, OSHA, and FDA policies are often in terms of promoting health and safety alone, without requirements that the cost be considered, and in some cases with requirements that cost not be considered.

While the legislative mandates are often structured quite narrowly, agencies may often interpret them in a more flexible manner. There may be some informal balancing of costs and benefits under different guises, such as a stated desire not to cause adverse effects on employment. The EPA pesticide regulation program strikes such a balance despite more uncompromising stated policy objectives, though Cropper et al. (1992) estimated that the cost per case of cancer avoided by this program had a value of $35 million, which is still quite high.

Suppose for concreteness we take the $7 million value per statistical life as the reasonable reference point for efficient levels of expenditure on saving statistical lives. Using that benchmark, the regulations that pass a benefit-cost test end with the radionuclides/uranium mines regulation, and the group of regulations that fail a benefit-cost test begins with the roadway worker protection regulations. There are many regulations on the table that cost in excess of $140 million per statistical life saved. The most expensive regulation listed is the EPA solid waste disposal facility criteria, which cost $100 billion per statistical life saved. The total budget for such regulations is of course less than $100 billion, as astronomical cost per life saved figures such as this emerge when dividing reasonably substantial cost amounts by very small mortality risk reductions.

The differences in the efficacy of government regulation are embodied in what Breyer (1993) popularized as the 90-10 Principle. This hypothesis is that society spends 90 percent of its costs to address the last 10 percent of the risk. Put somewhat differently, the first 10 percent of the costs will eliminate 90 percent of the risk. While this principle of decreasing marginal efficacy of regulatory costs is impressionistic and not based on empirical evidence, it is consistent with the widely differing cost-effectiveness of risk regulation policies. Thus, reallocations of expenditures toward policies at the top of Table 2 would save more lives at considerably less cost.[54]

From the standpoint of using society's resources in the most cost-effective way to save lives, uses of the funds in contexts other than those involving government regulation might provide a higher return in expected lives saved per dollar expended. Tengs

[54] Similarly, targeted allocations within programs would have the same benefits. Hamilton and Viscusi (1999) show for a large sample of Superfund sites that 10 percent of the expenditures would eliminate over 99 percent of the cancer risk.

et al. (1995) present an inventory of hundreds of estimates of costs per life saved from medical contexts. Analyses of these interventions may not be comparable to the regulatory assessments by Morrall (1986, 2003), who maintains consistent assumptions so that costs per life saved are comparable across regulations. Nevertheless, the Tengs et al. (1995) compilation highlights the often dramatic cost beneficial character of much medical care. Childhood immunizations, for example, are extremely valuable in terms of being a low cost mechanism for saving lives.

8. Risk-risk analysis

The existence of tradeoffs involving different kinds of risks frequently arises in regulatory situations. Perhaps the classic example is that involving FDA approvals for new prescription drugs, which focuses on the safety and efficacy of new drugs. The agency must balance two different kinds of health risks. FDA may reject a drug that is safe and effective, thus committing a Type I error. Alternatively, there is the risk that the agency will approve a drug that is not safe and effective and may in fact be dangerous, thus committing a Type II error. Note that both Type I errors and Type II errors impose health costs on the population. However, the risks differ in character. Type II risks are errors of commission and Type I errors are errors of omission. Because blame and adverse repercussions for the agency are generally higher for Type II errors, FDA places greater weight on these errors. From an economic efficiency standpoint, the risks should be treated symmetrically. The existence of such risk-risk tradeoffs is not restricted to FDA approvals but arises in a wide variety of regulatory situations.

8.1. Health effects

The existence of regulations that impose extremely large costs per life saved not only is a waste of economic resources, but also may harm individual health. A major empirical pattern that has been extremely well-documented is that greater affluence increases life expectancy and other measures of health status. Thus, to the extent that government regulations impose costs that divert funds away from expenditures that could have been made for health-enhancing commodities, such as better health care and safer products, pursuing such regulatory policies may have a counterproductive effect on overall health.

Some early estimates of this linkage focus on simple regression models linking income with mortality, yielding estimates that the level of income loss that would be sufficient to lead to the loss of one statistical life is under \$10 million. Results such as these were cited by Judge Stephen Williams in suggesting that OSHA regulations may be inefficient and may be counterproductive in terms of their effect on health.[55] Results

[55] See *International Union, UAW v. OSHA*, 938 F2d 1310 (DC Cir 1991) at 1326 (Williams concurring). The U.S. General Accounting Office (1992) also examined the policy implications of risk-risk analysis for OSHA rulemaking. Other early treatments of these issues include Keeney (1990); Graham, Hung-Chang, and Evans (1992); Lutter and Morrall (1994); and Viscusi (1994).

of such studies linking mortality to health status are complicated by the fact that the relationship is two-directional, as better health status also affects one's earnings. Moreover, the estimates of the income loss that would lead to the loss of one statistical life using this methodology were very close to estimates of the value of saving a statistical life, so that an expenditure of $7 million to save a statistical life would be almost entirely counterproductive in that this risk reduction expenditure would lead to a diversion of resources from a bundle of consumer expenditures that otherwise would have saved almost one statistical life.[56] Early estimates, such as those by Keeney (1990), tended to be low. A major difficulty has been controlling for the simultaneity of income and health. This was done by Chapman and Hariharan (1994).

The various approaches to estimating risk-risk tradeoffs can be reconciled by formalizing the theoretical linkage between the value of statistical life from the standpoint of saving lives as well as the level of income loss that leads to the loss of a statistical life. Even if the regulation does not pass a benefit-cost test, at the very minimum it should enhance safety or lead to $\Delta\pi > 0$ in terms of our earlier notation in which π is the probability of survival. The components of $\Delta\pi$ consist of the marginal effect of regulation-induced safety level effects Δs on the value of π, the marginal effect of Δs in decreasing health investments, and the effect of the financial cost Δy of regulation on risk due to the positive income elasticity of demand for health-related expenditures. The condition that regulations promote safety is that

$$\Delta\pi = \frac{\partial\pi}{\partial s}\Delta s + \frac{\partial\pi\,\partial h}{\partial h\partial s}\Delta s + \frac{\partial\pi\,\partial h}{\partial h\partial y}\Delta y > 0. \tag{26}$$

After some manipulation of equation (26), it can be shown that government regulations will have a net risk-reducing effect on the risk level if

$$\begin{aligned}&\text{Average Regulatory Cost per Life Saved}\\ &\quad < (\text{Marginal Value of Life})/(\text{ Marginal Propensity to Spend on Health}).\end{aligned} \tag{27}$$

The key concern is consequently what the magnitude of the term on the right side of this inequality is. Clearly the average regulatory cost per life saved that could be expended before regulations become counterproductive is quite closely linked to the value of statistical life. The critical cutoff for cost-effective regulations is the value of statistical life divided by the marginal propensity to spend on health. If the calculation includes only favorable health expenditures that are mortality-reducing in calculations pertaining to this denominator, as in Viscusi (1994, 1998), the level of expenditure of regulations before these efforts become counterproductive is estimated to be $70 million if the value of statistical life is $7 million. However, if this calculation also recognizes

[56] These estimates for the U.S. are similar in magnitude to those found for Sweden by Gerdtham and Johanneson (2002), whose risk-risk analysis estimated that the income loss that would lead to the loss of one statistical life ranged from $6.8 million to $9.8 million.

the income relatedness of expenditures that harm individual health, such as those pertaining to smoking, drinking, and lack of exercise, as in Lutter, Morrall, and Viscusi (1999), then the level of expenditures before regulations become counterproductive is a cutoff level under $30 million. Thus, for this cutoff level, all regulations at or below the asbestos/construction regulation appearing in Table 2 combine a low level of expenditure with such a modest direct safety enhancing effect that on balance these regulations are harming individual health rather than benefiting health. Considering a set of regulations similar to those in Table 2, Hahn, Lutter, and Viscusi (2000) present a set of estimates of the net effects of regulations on risk levels that reflect the risk-risk tradeoffs involved.

The use of protective equipment, such as seatbelts, reduces the risk to the wearer, thus providing potentially valuable private benefits. Blomquist (1979) uses evidence from the tradeoff users of seatbelts made between disutility (time and discomfort) and safety to derive an estimate of VSL in this context that is not unlike labor market estimates. From a policy standpoint, Arnould and Grabowski (1981) estimate that the private risk reduction benefits from seatbelts offset the costs so that failure to buckle up represents a market failure. Indeed, recent estimates by Levitt and Porter (2001) indicate that seatbelts save lives at a cost of $30,000 per statistical life as compared to $1.8 million for air bags.

Such estimates take into account only the risk reduction aspect of seatbelts and ignore potentially offsetting behavioral consequences. Safety regulations may also not have their intended effects if people reduce their precautionary behavior in response to the regulation. Thus, there may be a moral hazard problem with respect to regulations that enhance the safety level associated with any particular level of precautionary behavior.

Two different mechanisms for such a response have been distinguished in the literature. In his analysis of individual responses to seatbelt regulations, Peltzman (1975) hypothesized that the effect of seatbelts in reducing the risk level would be to diminish the level of care taken by drivers, thus dampening and possibly eliminating the safety reduction benefits of seatbelts. In Viscusi (1984, 1992) the hypothetical mechanism influencing the performance of safety caps stemmed from a misperception of the extent to which safety devices were protective. Thus, if people believe that safety caps on medicine made these products childproof rather than simply more difficult to open, they would become more lax about the extent of safety precautions than they would be had they been aware of the actual level of risk reduction produced by safety caps.

We can illustrate these relationships with a simple model. Let f be the stringency of the government policy as reflected in the technological safety-related requirements. The other determinant of the accident risk is the individual's safety-related effort e. Thus, we have an overall accident risk probability function $p(e, f)$, where higher levels of both e and f reduce the accident risk but at a diminishing rate. If an injury occurs, the accident imposes a monetary equivalent loss of z. Taking safety precautions does, however, impose a cost on the individual given by $v(e)$, where both v', $v'' > 0$. The individual in this model, in which effort and harm have monetary equivalents, has an income level of y. Suppose that the government's choice of the regulatory stringency f

is given, and that the individual chooses the level of safety-related effort e to

$$\operatorname*{Max}_{e} Z = [1 - p(e, f)][I - v(e)] + p(e, f)[I - v(e) - z], \tag{28}$$

or

$$\operatorname*{Max}_{e} Z = I - v(e) - p(e, f)z, \tag{29}$$

leading to the first-order condition

$$-p_e z = v_e. \tag{30}$$

The marginal reduction of the loss due to extra safety-related effort is equated to the marginal disutility of effort.

If the stringency of government regulation is increased, the safety-related effort will decline, or

$$\frac{de}{df} = \frac{-p_{ef} z}{p_{ee} z + v_{ee}} < 0, \tag{31}$$

if we assume that $p_{ef} > 0$ as in previous analyses of this issue.

The extent of the behavioral response is an empirical issue. On a theoretical basis the level of precautions should decrease, but whether the offsetting behavior will lead risk levels to be greater than would prevail had there been no regulation can only occur if one imposes strong restrictions on functional forms. If, however, people overestimate the efficacy of the regulatory policy s, then there is greater potential for a counterproductive effect.

Empirical evidence for the consequence of seatbelt regulation has not been conclusive. The effect of seatbelt regulations has been controversial, as Peltzman (1975) found that there was no evidence that the advent of seatbelts had a beneficial effect on safety. Similarly, a comprehensive review of the evidence and new empirical work by Blomquist (1988, 1991) found that use of seatbelts led to increased deaths among motorcyclists and pedestrians, which is consistent with the behavioral response model in which drivers increase their driving speed if they are protected by seatbelts. Keeler (1994) found that urban speed limits reduced fatalities, but rural speed limits did not, which he hypothesized might be the result of offsetting behavior. More recent evidence by Cohen and Einav (2003), however, failed to find evidence of the offsetting behavior.[57]

The empirical evidence regarding the effect of safety caps on child poisoning is more clear-cut, largely because it derives from a situation in which people erroneously believe that the safety devices are more effective than they actually are. The majority of aspirin child poisonings have been from safety-capped bottles even to a greater extent than would be predicted by the safety cap share of aspirin sales. Related products, such as acetaminophen, exhibited an increase in risk levels once safety caps were introduced as consumers apparently became more lax about the need to be more protective about medicines generally.

[57] For additional evidence along these lines that is more supportive of seatbelts, see Crandall et al. (1986).

9. The enforcement and performance of government regulations

The efficacy of government regulations in enhancing health and safety is reflected to a large extent in the cost-effectiveness statistics presented in Table 2. There have been considerable expenditures made that were expected to have modest effects on risk reduction under three key assumptions. First, the analyses assume that there will be no offsetting behavioral responses that diminish the assumed efficacy of the health and safety standards. Thus, drivers for example do not become riskier in their driving habits if they use seatbelts. Second, the risk-risk effects that arise from the fact that regulatory expenditures divert our resources from other health enhancing expenditures are not taken into account. Third, the analyses assume that there will be full compliance with the regulatory standard. With regulations that have substantial costs of compliance, it is by no means ensured that firms will make the requisite expenditures to achieve compliance.

Because enforcement efforts are costly, it is not generally socially optimal to have enforcement efforts that induce full compliance. The economic theory of enforcement began with the analysis of criminal behavior by Becker (1968) and continued with a series of papers by Polinsky and Shavell (1979, 1992, 1994) and others. As an illustration of the pertinence of the economic analysis of enforcement to regulatory contexts, consider the work by Cohen (1987, 1992), who develops models of optimal enforcement activity that demonstrate the tradeoffs involved for environmental policy. The social welfare maximizing government agency in the Cohen (1987) model minimizes total social costs, which include environmental damages, cleanup or recovery costs, private-resource costs, preventive expenditures, monitoring expenses, and detection expenses. The optimizing regulator will set the penalty function to induce the socially optimal level of safety. With perfect enforcement, the expected penalty equals the expected environmental damage plus cleanup cost. However, the probability of detection is not 1.0 because monitoring and detection will not be perfect given the costs associated with enforcement, so that the penalty should be increased accordingly to provide appropriate incentives.

In practice, non-complying firms often face very low regulatory penalties coupled with a small probability of detection. To examine the firm's incentives for compliance, consider a model for analyzing whether firms comply with regulations follows the same general approach as economic models of crime. Risk-neutral firms will choose to comply with the regulation if the costs of compliance are less than the expected costs of noncompliance, or

$$\begin{aligned}&\text{Costs of Compliance} < \text{Probability of Inspection}\\&\quad\times \text{Number of Violations per Inspection} \times \text{Penalty per Violation}.\end{aligned}\tag{32}$$

Whether firms find it desirable to comply with a regulation depends not just on compliance costs, which tend to be high, but also on the regulatory enforcement approach, which varies considerably by agency. The Occupational Safety and Health Administration (OSHA) has received the most attention for inadequate compliance with its regulations as the probability of inspection for any given firm is under 1/100 annually. If the firm is inspected, the inspectors generally identify a few violations and impose penalties that are dwarfed by market incentives for safety. The inspection strategy for EPA water pollution regulations requires annual inspections as well as the filing of monthly discharge monitoring reports, leading to much greater incentives for compliance. In the case of mass marketed consumer products, such as lawn mowers and prescription drugs, compliance is readily monitorable.

The performance of regulatory policies has varied substantially due to differences in the efficacy of the different approaches as well as differences in enforcement. Several studies have estimated effects of the broadly based OSHA inspections on safety, which were found to be zero or small in most studies.[58] When these inspections are targeted strategically by the agency, the performance may fare better. Inspections that are undertaken only after the agency inspects the on-site incident records have led to lower reported injury rate levels.[59] EPA water pollution standards have had a much better track record, as has the EPA phasedown of lead in gasoline. The lead phasedown regulation is particularly noteworthy in that it is a success story in terms of having costs below the economic benefits. In other instances, such as seatbelt regulations and safety caps, the behavioral response to regulations has muted or offset the beneficial effect of the safety-enhancing technology.

10. Conclusion

Economics plays a central role in the analysis and development of risk and environmental regulations. The sources of market failure in this arena provide the impetus for an active role of regulatory policy. The economics literature on informational failures, externalities, and irrational choice often use risk and environmental contexts as the paradigmatic examples of these market failures.

An economic analysis of the regulatory policy is a required component of the policy development process in the U.S., and preparation of regulatory impact analyses for prospective regulations is increasingly becoming the norm throughout the world. While economic assessments of benefits and costs may inform regulatory policy decisions, the restrictive legislative mandates of regulatory agencies often require that the agency adopt objectives other than a comprehensive balancing of benefits and costs.

Economic analysis of regulations has proven to be important both before and after the regulations are issued. The regulatory impact analyses of prospective regulations provide an evaluation of regulatory costs and benefits that serves as the main basis

[58] See Viscusi (1992) and Smith (1979b), among others, who show zero or small effects of OSHA inspections. Scholz and Gray (1990) report greater effects as, for example, there will be a one percent reduction in injury risks for a ten percent increase in enforcement.

[59] Ruser and Smith (1988) provide a detailed analysis of these records-check inspections and their effect on published injury rates.

for regulatory policy debates and internal administrative review. Analyses of the comparative cost-effectiveness of regulatory policies have highlighted possible regulatory imbalances and have indicated how benefits to society could be increased at less cost. *Ex post* economic analyses of regulatory performance have served a variety of functions, such as highlighting potential inadequacies in enforcement, as well as providing empirical evidence of the extent of behavioral response to regulatory policies.

Often the stated policy concerns are not with respect to an economic efficiency objective but a commitment to improve the environment, make workplaces safe, or promote the safety and efficacy of prescription drugs. Even if the concern is with risk reduction alone, economic analysis provides numerous insights. The performance of the regulatory policy will hinge on the incentives these policies create, both with respect to incentives for regulatory compliance as well as incentives for individuals to undertake risky behaviors that may diminish regulatory effectiveness. Economic studies of risk-risk tradeoffs have also demonstrated that there may be important health-related opportunity costs that may be consequential even if the concern is with risk reduction irrespective of cost.

Acknowledgements

This research is supported by the Harvard Olin Center for Law, Economics and Business. Rob Stavins provided excellent detailed comments on a draft of this paper.

References

Adler, R., Pittle, D. (1984). “Cajolery and Command: Are Education Campaigns an Adequate Substitute for Regulation?” Yale Journal of Regulation 2, 159–194.

Aldy, J., Viscusi, W.K. (2004). “Age Variations in Workers’ Value of Statistical Life”. NBER Working Paper, No. w10199. Internet: http://www.nber.org/papers/w10199.

American Law Institute (1991). Enterprise Responsibility for Personal Injury—Reporters’ Study. American Lung Institute, Philadelphia.

Ames, B.N., Gold, L.S. (1990). “Chemical Carcinogenesis: Too Many Rodent Carcinogens”. Proceedings of the National Academy of Sciences, USA 87, 7772–7776.

Ames, B.N., Profet, M., Gold, L.S. (1990). “Nature’s Chemicals and Synthetic Chemicals: Comparative Toxicology”. Proceedings of the National Academy of Sciences, USA 87, 7777–7781.

Arnould, R.J., Grabowski, H. (1981). “Auto Safety Regulation: An Analysis of Market Failure”. Bell Journal of Economics 12 (1), 27–48.

Atkinson, S.E., Halvorsen, R. (1990). “The Valuation of Risks to Life: Evidence from the Market for Automobiles”. Review of Economics and Statistics 72 (1), 133–136.

Becker, G.S. (1968). “Crime and Punishment: An Economic Approach”. Journal of Political Economy 76 (2), 169–217.

Becker, G.S., Grossman, M., Murphy, K.M. (1994). “An Empirical Analysis of Cigarette Addiction”. American Economic Review 84, 396–418.

Becker, G.S., Murphy, K.M. (1988). “A Theory of Rational Addiction”. Journal of Political Economy 96, 675–700.

Bernheim, B.D., Rangel, A. (2004). “Addiction and Cue-Triggered Decision Processes”. American Economic Review 94 (5), 1558–1590.

Bettman, J.R., Payne, J.W., Staelin, R. (1987). “Cognitive Considerations in Presenting Risk Information”. In: Viscusi, W.K., Magat, W.A. (Eds.), Learning about Risk: Consumer and Worker Responses to Hazard Information. Harvard University Press, Cambridge, MA.

Blomquist, G.C. (1979). “Value of Life Saving: Implications of Consumption Activity”. Journal of Political Economy 87 (3), 540–558.

Blomquist, G.C. (1988). The Regulation of Motor Vehicle and Traffic Safety. Kluwer Academic Publishers, Boston.

Blomquist, G.C. (1991). “Motorist Use of Safety Equipment: Expected Benefits or Risk Incompetence”. Journal of Risk and Uncertainty 4 (2), 135–152.

Breyer, S. (1993). Breaking the Vicious Circle: Toward Effective Risk Regulation. Harvard University Press, Cambridge, MA.

Brookshire, D.S., Thayer, M.A., Tschirhart, J., Schulze, W.D. (1985). “A Test of the Expected Utility Model: Evidence from Earthquake Risks”. Journal of Political Economy 93 (2), 369–389.

Burmaster, D.E., Harris, R.N. (1993). “The Magnitude of Compounding Conservatism in Superfund Risk Assessments”. Risk Analysis 13 (2), 131–134.

Camerer, C.F., Kunreuther, H. (1989). “Decision Processes for Low Probability Events: Policy Implications”. Journal of Policy Analysis and Management 8 (4), 565–592.

Camerer, C.F., Weber, M. (1992). “Recent Developments in Modeling Preferences: Uncertainty and Ambiguity”. Journal of Risk and Uncertainty 5, 325–370.

Chapman, K.S., Hariharan, G. (1994). “Controlling for Causality in the Link from Income to Mortality”. Journal of Risk and Uncertainty 8 (1), 85–93.

Chemical Manufacturers Association, (1991). “Analysis of the Impact of Exposure Assumptions on Risk Assessment of Chemicals in the Environment”. Prepared by Risk Focus, Versar, Inc.

Cohen, A., Einav, L. (2003). “The Effects of Mandatory Seat Belt Laws on Driving Behavior and Traffic Fatalities”. The Review of Economics and Statistics 85 (4), 828–843.

Cohen, M.A. (1986). “The Costs and Benefits of Oil Spill Prevention and Enforcement”. Journal of Environmental Economics and Management 13, 167–188.

Cohen, M.A. (1987). “Optimal Enforcement Strategy to Prevent Oil Spills: An Application of a Principal-Agent Model with Moral Hazard”. Journal of Law and Economics 30, 23–51.

Cohen, M.A. (1992). “Environmental Crime and Punishment: Legal/Economic Theory and Empirical Evidence of Enforcement of Federal Environmental Statutes”. The Journal of Criminal Law and Criminology 82 (4), 1054–1108.

Crandall, R.W., Gruenspecht, H.K., Keeler, T.E., Lave, L.B. (1986). Regulating the Automobile. Brookings Institute, Washington.

Cropper, M.L., Evans, W.N., Berardi, S.J., Ducla-Soares, M.M., Portney, P.R. (1992). “The Determinants of Pesticide Regulation: A Statistical Analysis of EPA Decision Making”. Journal of Political Economy 100 (1), 175–197.

Cropper, M.L., Portney, P.R. (1990). “Discounting and the Evaluation of Lifesaving Programs”. Journal of Risk and Uncertainty 3 (4), 369–379.

Cullen, A.C. (1994). “Measures of Compounding Conservatism in Probabilistic Risk Assessment”. Risk Analysis 14 (4), 389–393.

Desvousges, W., et al. (1993). “Measuring Natural Resource Damages with Contingent Valuation: Tests of Validity and Reliability”. In: Hausman, J. (Ed.), Contingent Valuation: A Critical Assessment. North Holland Press, Amsterdam, pp. 91–164.

Diamond, P.A., Hausman, J.A. (1994). “Contingent Valuation: Is Some Number Better than No Number?” Journal of Economic Perspectives 8 (4), 45–64.

Ellsberg, D. (1961). “Risk, Ambiguity, and the Savage Axioms”. Quarterly Journal of Economics 75, 643–669.

European Commission (1996). Technical Guidance Document in Support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances and Commission Regulation No. 1488/94 on

Risk Assessment for Existing Substances. Office for Official Publications of the European Communities, Luxembourg.

Finkel, A.M. (1989). "Is Risk Assessment Really Too Conservative?". Columbia Journal of Environmental Law 14, 427–467.

Finley, B., Paustenbach, D. (1994). "The Benefits of Probabilistic Exposure Assessment: Three Case Studies Involving Contaminated Air, Water, and Soil". Risk Analysis 14 (1), 53–73.

Fischhoff, B., Lichtenstein, S., Slovic, P., Derby, S.L., Keeney, R. (1981). Acceptable Risk. Cambridge University Press, Cambridge, U.K.

Freeman, A.M. III (1993). The Measurement of Environmental and Resource Values. Resources for the Future, Washington.

Freeman, A.M. III (2002). "Environmental Policy Since Earth Day I: What Have We Gained?". The Journal of Economic Perspectives 16 (1), 125–146.

Gayer, T. (2004). "The Fatality Risks of Sport-Utility Vehicles, Vans, and Pickups Relative to Cars". Journal of Risk and Uncertainty 28 (2), 103–133.

Gerdtham, U.-G., Johanneson, M. (2002). "Do Life-Saving Regulations Save Lives?". Journal of Risk and Uncertainty 24 (3), 231–249.

Gollier, C., Treich, N. (2003). "Decision-Making Under Scientific Uncertainty: The Economics of the Precautionary Principle". Journal of Risk and Uncertainty 27 (1), 77–103.

Graham, J., Hung-Chang, B., Evans, J.S. (1992). "Poorer is Riskier". Risk Analysis 12 (3), 333–337.

Gruber, J., Köszegi, B. (2001). "Is Addiction 'Rational'? Theory and Evidence". Quarterly Journal of Economics 116 (4), 1261–1303.

Hahn, R.W., Lutter, R.W., Viscusi, W.K. (2000). Do Federal Regulations Reduce Mortality? AEI-Brookings Joint Center for Regulatory Studies, Washington.

Hanemann, W.M. (1994). "Valuing the Environment Through Contingent Valuation". Journal of Economic Perspectives 8 (4), 19–43.

Hamilton, J.T. (1995). "Testing for Environmental Racism: Prejudice, Profits, Political Power?" Journal of Policy Analysis and Management 14 (1), 107–132.

Hamilton, J.T., Viscusi, W.K. (1999). Calculating Risks?: The Spatial and Political Dimensions of Hazardous Waste Policy. MIT Press, Cambridge, MA.

Harrington, W., Portney, P.R. (1987). "Valuing the Benefits of Health and Safety Regulations". Journal of Urban Economics 22 (1), 101–112.

Hersch, J. (1998). "Compensating Differentials for Gender-Specific Job Injury Risks". American Economic Review 88 (3), 598–607.

Hersch, J. (2002). "Breast Implants: Regulation, Litigation, and Science". In: Viscusi, W.K. (Ed.), Regulation through Litigation. AEI-Brookings Joint Center for Regulatory Studies, Washington, pp. 142–177.

Hersch, J., Pickton, T.S. (1995). "Risk-Taking Activities and Heterogeneity of Job-Risk Tradeoffs". Journal of Risk and Uncertainty 11, 205–217.

Hersch, J., Viscusi, W.K. (1990). "Cigarette Smoking, Seatbelt Use, and Differences in Wage-Risk Tradeoffs". Journal of Human Resources 25 (2), 202–227.

Horowitz, J.K., Carson, R.T. (1990). "Discounting Statistical Lives". Journal of Risk and Uncertainty 3 (4), 403–413.

Huber, P. (1988). Liability: The Legal Revolution and its Consequences. Basic Books, New York.

Ippolito, P.M., Ippolito, R.A. (1984). "Measuring the Value of Life Saving from Consumer Reactions to New Information". Journal of Public Economics 25, 53–81.

Jenni, K.E., Loewenstein, G. (1997). "Explaining the 'Identifiable Victim Effect' ". Journal of Risk and Uncertainty 14 (3), 235–257.

Johannesson, M., Johansson, P.-O., O'Conor, R.M. (1996). "The Value of Private Safety Versus the Value of Public Safety". Journal of Risk and Uncertainty 13 (3), 263–275.

Johansson, P.-O. (2002). "The Definition and Age-Dependency of the Value of a Statistical Life". Journal of Risk and Uncertainty 25 (3), 251–263.

Jones-Lee, M.W. (1976). The Value of Life: An Economic Analysis. University of Chicago Press, Chicago.

Jones-Lee, M.W. (1989). The Economics of Safety and Physical Risk. Basil Blackwell, Oxford.

Jones-Lee, M.W. (1991). "Altruism and the Value of Other People's Safety". Journal of Risk and Uncertainty 4 (2), 213–219.

Jones-Lee, M.W., Hammerton, M., Philips, P.R. (1985). "The Value of Safety: Results of a National Sample Survey". Economic Journal 95, 49–72.

Kahneman, D., Knetsch, J.L. (1992). "Valuing Public Goods: The Purchase of Moral Satisfaction". Journal of Environmental Economics and Management 22, 57–70.

Keeler, T.E. (1994). "Highway Safety, Economic Behavior, and Driving Environment". American Economic Review 84 (3), 684–693.

Keeney, R.L. (1990). "Mortality Risks Induced by Economic Expenditures". Risk Analysis 10 (1), 147–159.

Kniesner, T.J., Leeth, J.D. (1991). "Compensating Wage Differentials for Fatal Injury Risk in Australia, Japan, and the United States". Journal of Risk and Uncertainty 4 (1), 75–90.

Kniesner, T.J., Viscusi, W.K., Ziliak, J.P. (2006). "Life-Cycle Consumption and the Age-Adjusted Value of Life". Contributions to Economic Analysis and Policy 5 (1), Article 4.

Krupnick, A., Alberini, A., Cropper, M., Simon, N., O'Brien, B., Goeree, R., Heintzelman, M. (2002). "Age, Health, and the Willingness to Pay for Mortality Risk Reductions: The Contingent Valuation Survey of Ontario Residents". Journal of Risk and Uncertainty 24 (2), 161–186.

Kunreuther, H., et al. (1978). Disaster Insurance Protection: Public Policy Lessons. Wiley, New York.

Kunreuther, H., Hogarth, R., Meszaros, J. (1993). "Insurer Ambiguity and Market Failure". Journal of Risk and Uncertainty 7 (1), 71–87.

Levitt, S.D., Porter, J. (2001). "Sample Selection in the Estimation of Air Bag and Seat Belt Effectiveness". Review of Economics and Statistics 83 (4), 603–615.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, U., Combs, B. (1978). "Judged Frequency of Lethal Events". Journal of Experimental Psychology: Human Learning and Memory 4, 551–578.

Liu, J.-T., Hsieh, C.-R. (1995). "Risk Perception and Smoking Behavior: Empirical Evidence from Taiwan". Journal of Risk and Uncertainty 11 (2), 139–157.

Löfstedt, R. (2004). "The Swing of the Regulatory Pendulum in Europe: From Precautionary Principle to (Regulatory) Impact Analysis". Journal of Risk and Uncertainty 28 (3), 237–260.

Lutter, R., Mader, E. (2002). "Litigating Lead-Based Paint Hazards". In: Viscusi, W.K. (Ed.), Regulation through Litigation. AEI-Brookings Joint Center for Regulatory Studies, Washington, pp. 106–135.

Lutter, R., Morrall, J.F. III (1994). "Health-Health Analysis: A New Way to Evaluate Health and Safety Regulation". Journal of Risk and Uncertainty 8 (1), 43–46.

Lutter, R., Morrall, J.F. III, Viscusi, W.K. (1999). "The Cost-per-Life-Saved Cutoff for Safety-Enhancing Regulations". Economic Inquiry 37 (4), 599–608.

Magat, W.A., Viscusi, W.K. (1992). Informational Approaches to Regulation. MIT Press, Cambridge, MA.

Magat, W.A., Viscusi, W.K., Huber, J. (1996). "A Reference Lottery Metric for Valuing Health". Management Science 42 (8), 1118–1129.

Manning, R.L. (1994). "Changing Rules in Tort Law and the Market for Childhood Vaccines". Journal of Law and Economics 37 (1), 247–275.

Manning, W.G., Keeler, E.B., Newhouse, J.P., Sloss, E.M., Wasserman, J. (1989). "Taxes of Sin: Do Smokers and Drinkers Pay Their Way?" Journal of the American Medical Association 26 (11), 1604–1609.

McClelland, G.H., Schulze, W.D., Hurd, B. (1990). "The Effects of Risk Beliefs on Property Values: A Case Study of a Hazardous Waste Site". Risk Analysis 10 (4), 485–497.

Michigan Manufacturers Association (1993). "A Comparison of Monte Carlo Simulation-Based Exposure Estimates with Estimates Calculated Using EPA and Suggested Michigan Manufacturers Association Exposure Factors". Prepared by ENVIRON Corporation.

Mishan, E.J. (1971). "Evaluation of Life and Limb: A Theoretical Approach". Journal of Political Economy 79 (4), 687–705.

Moore, M.J., Viscusi, W.K. (1990). Compensating Mechanisms for Job Risks: Wages, Workers' Compensation, and Product Liability. Princeton University Press, Princeton.

Morrall, J.F. III (1986). "A Review of the Record". Regulation 10 (2), 25–34.

Morrall, J.F. III (2003). "Saving Lives: A Review of the Record". Journal of Risk and Uncertainty 27 (3), 221–237.

Murray, C.J.L., Lopez, A.D. (1996). The Global Burden of Disease: Summary. World Health Organization, Geneva.

Nader, R. (1965). Unsafe at Any Speed. Grossman Publishers, New York.

National Academy of Sciences (1990). Developing New Contraceptives: Obstacles and Opportunities. National Academy Press, Washington.

National Research Council (1983). Risk Assessment in the Federal Government: Managing the Process. National Academy Press, Washington.

National Research Council (1994). Science and Judgment in Risk Assessment. National Academy Press, Washington.

National Research Council (1996). Carcinogens and Anticarcinogens in the Human Diet. National Academy Press, Washington.

National Safety Council (2004). Injury Facts. National Safety Council, Itasca, IL.

Nichols, A.L., Zeckhauser, R.J. (1986). "The Perils of Prudence: How Conservative Risk Assessments Distort Regulation". Regulation 10 (2), 13–24.

Office of Science and Technology Policy, Executive Office of the President (1985). Chemical Carcinogens: A Review of the Science and its Associated Principles. Federal Register 50: 10371–10442, March 14.

Peltzman, S. (1975). "The Effects of Automobile Safety Regulation". Journal of Political Economy 83 (4), 677–725.

Polinsky, A.M., Shavell, S. (1979). "The Optimal Tradeoff between the Probability and Magnitude of Fines". American Economic Review 69 (5), 880–891.

Polinsky, A.M., Shavell, S. (1992). "Enforcement Costs and the Optimal Magnitude and Probability of Fines". Journal of Law and Economics 35 (1), 133–148.

Polinsky, A.M., Shavell, S. (1994). "Should Liability Be Based on the Harm to the Victim or the Gain to the Injurer?" Journal of Law, Economics, and Organization 10 (2), 427–437.

Portney, P.R. (1981). "Housing Prices, Health Effects, and Valuing Reductions in Risk of Death". Journal of Environmental Economics and Management 8, 72–78.

Portney, P.R. (1994). "The Contingent Valuation Debate: Why Economists Should Care". Journal of Economic Perspectives 8 (4), 3–17.

Posner, R.A. (2004). Catastrophe: Risk and Response. Oxford University Press, New York.

Raiffa, H. (1961). "Risk, Ambiguity, and the Savage Axioms: Comment". Quarterly Journal of Economics 75, 690–694.

Revesz, R.L. (1999). "Environmental Regulation, Cost-Benefit Analysis, and the Discounting of Human Lives". Columbia Law Review 99, 941–1017.

Rosen, S. (1988). "The Value of Changes in Life Expectancy". Journal of Risk and Uncertainty 1 (3), 285–304.

Ruser, J.W., Smith, R.S. (1988). "The Effect of OSHA Records-Check Inspections on Reported Occupational Injuries in Manufacturing Establishments". Journal of Risk and Uncertainty 1 (4), 415–435.

Samuelson, W., Zeckhauser, R. (1988). "Status Quo Bias in Decision Making". Journal of Risk and Uncertainty 1 (1), 7–59.

Schelling, T.C. (1968). "The Life You Save May Be Your Own". In: Chase, S.B. Jr. (Ed.), Problems in Public Expenditure Analysis. Brookings Institute, Washington, pp. 127–162.

Schelling, T.C. (1978). Micromotives and Macrobehavior. W.W. Norton, New York.

Schelling, T.C. (1984). Choice and Consequence. Harvard University Press, Cambridge, MA.

Scholz, J.T., Gray, W.B. (1990). "OSHA Enforcement and Workplace Injuries: A Behavioral Approach to Risk Assessment". Journal of Risk and Uncertainty 3 (3), 283–305.

Shavell, S. (1987). Economic Analysis of Accident Law. Harvard University Press, Cambridge, MA.

Shavell, S. (2004). Foundations of Economic Analysis of Law. Harvard University Press, Cambridge, MA.

Shepard, D.S., Zeckhauser, R.J. (1984). "Survival Versus Consumption". Management Science 30 (4), 423–439.

Smith, R.S. (1979a). "Compensating Wage Differentials and Public Policy: A Review". Industrial and Labor Relations Review 32 (3), 339–352.

Smith, R.S. (1979b). "The Impact of OSHA Inspections on Manufacturing Injury Rates". Journal of Human Resources 14, 145–170.

Smith, V.K., Desvousges, W.H. (1986). "The Value of Avoiding a LULU: Hazardous Waste Disposal Sites". Review of Economics and Statistics 68, 293–299.

Sunstein, C. (2002). Risk and Reason: Safety, Law, and the Environment. Cambridge University Press, Cambridge, U.K.; New York.

Tengs, T.O., Adams, M.E., Pliskin, J.S., Safran, D.G., Siegel, J.E., Weinsten, M.C., Graham, J.D. (1995). "Five-Hundred Life-Saving Interventions and their Cost-Effectiveness". Risk Analysis 15 (3), 369–390.

Thaler, R., Rosen, S. (1976). "The Value of Saving a Life: Evidence from the Labor Market". In: Terleckyj, N. (Ed.), Household Production and Consumption. National Bureau of Economic Research, New York, pp. 265–298.

U.S. Environmental Protection Agency (1987). The Risk Assessment Guidelines of 1986. EPA-600/8-87/045. Guidelines for Carcinogen Risk Assessment; Guidelines for Mutagenicity Risk Assessment; Guidelines for the Health Risk Assessment of Chemical Mixtures; Guidelines for the Health Assessment of Suspect Developmental Toxicants; Guidelines for Estimating Exposures. U.S. EPA, Office of Health and Environmental Assessment, Washington.

U.S. Environmental Protection Agency (1988). Proposed Guidelines for Exposure-Related Measurements. Federal Register 53: 48830–48853, December 2.

U.S. Environmental Protection Agency (1992). Guidelines for Exposure Assessment. Federal Register 57: 22888–22938, May 29.

U.S. Environmental Protection Agency (1996). Proposed Guidelines for Carcinogen Risk Assessment. Federal Register 61: 17960-18011, April 23.

U.S. Environmental Protection Agency (2002). Technical Addendum: Methodologies for the Benefit Analysis of the Clear Skies Initiative. USEPA, Washington.

U.S. Environmental Protection Agency (2003). Technical Addendum: Methodologies for the Benefit Analysis of the Clear Skies Act of 2003. USEPA, Washington.

U.S. General Accounting Office (1992). Risk-Risk Analysis: OMB's Review of a Proposed OSHA Rule. Report to the Chairman, Committee on Governmental Affairs, U.S. Senate, July.

Viscusi, W.K. (1984). "The Lulling Effect: The Impact of Child-Resistant Packaging on Aspirin and Analgesic Ingestions". American Economic Review 74 (2), 324–327.

Viscusi, W.K. (1992). Fatal Tradeoffs: Public and Private Responsibilities for Risk. Oxford University Press, Oxford.

Viscusi, W.K. (1994). "Mortality Effects of Regulatory Costs and Policy Evaluation Criteria". Rand Journal of Economics 25 (1), 94–109.

Viscusi, W.K. (1997). "Alarmist Decisions with Divergent Risk Information". The Economic Journal 107, 1657–1670.

Viscusi, W.K. (1998). Rational Risk Policy. Oxford University Press, Oxford.

Viscusi, W.K. (2002). Smoke-Filled Rooms: A Postmortem on the Tobacco Deal. Chicago University Press, Chicago.

Viscusi, W.K., Aldy, J.E. (2003). "The Value of a Statistical Life: A Critical Review of Market Estimates Throughout the World". Journal of Risk and Uncertainty 27 (1), 5–76.

Viscusi, W.K., Hamilton, J.T., Dockins, P.C. (1997). "Conservative versus Mean Risk Assessments: Implications for Superfund Policies". Journal of Environmental Economics and Management 34, 187–206.

Viscusi, W.K., Hersch, J. (2001). "Cigarette Smokers as Job Risk Takers". Review of Economics and Statistics 83 (2), 269–280.

Viscusi, W.K., Magat, W.A. (1987). Learning about Risk: Consumer and Worker Responses to Hazard Information. Harvard University Press, Cambridge, MA.

Viscusi, W.K., Magat, W.A., Huber, J. (1987). "An Investigation of the Rationality of Consumer Valuations of Multiple Risks". Rand Journal of Economics 18 (4), 465–479.

Viscusi, W.K., Moore, M.J. (1993). “Product Liability, Research and Development, and Innovation”. Journal of Political Economy 101, 161–184.

Viscusi, W.K., Zeckhauser, R. (2004). “The Denominator Blindness Effect: Accident Frequencies and the Misjudgment of Recklessness”. American Law and Economics Review 6 (1), 72–94.

Zeckhauser, R. (1975). “Procedures for Valuing Lives”. Public Policy 23 (4), 419–464.

This page intentionally left blank

Chapter 10

TAXATION

LOUIS KAPLOW

School of Law, Harvard University, and National Bureau of Economic Research

Contents

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)01010-9

Abstract

This *Handbook* entry presents a conceptual, normative overview of the subject of taxation. It emphasizes the relationships among the main functions of taxation—notably, raising revenue, redistributing income, and correcting externalities—and the mapping between these functions and various forms of taxation. Different types of taxation as well as expenditures on transfers and public goods are each integrated into a common optimal tax framework with the income tax and commodity taxes at the core. Additional topics addressed include a range of dynamic issues, the unit of taxation, tax administration and enforcement, and tax equity.

Keywords

taxation, optimal taxation, income tax, redistribution, commodity tax, Ramsey taxation, taxation of capital, consumption tax, corporate tax, transfer tax, estate and gift tax, social security, state and local tax, international tax, transfers, welfare, public goods, cost-benefit analysis, benefit taxation, corrective taxation, environmental taxation, taxation and risk, transitions, capital gains, human capital, lifetime horizon, generational policy, taxation of family, tax administration, tax avoidance and evasion, tax enforcement, tax base, VAT, sales tax, tax equity, horizontal equity, inequality, poverty, progressivity

JEL classification: D61, D62, D63, H00, H20, H21, H23, H24, H26, H41, H43, H53, H55, H71, H87, I38, K34

1. Introduction

The subject of taxation is vast and has been a major focus of numerous economists over the ages. Accordingly, a single survey must be highly selective. Because there exists a four-volume *Handbook of Public Economics* (Auerbach and Feldstein, 1985, 1987, 2002a, 2002b), a substantial portion of which is devoted to taxation, and many other survey articles on various aspects of taxation, this review does not attempt to cover all the traditional topics, which would be impossible in any event. Instead, it aims to offer a guide that will complement existing work.

Specifically, this essay presents a conceptual, normative overview of the subject of taxation.[1] It emphasizes the relationships among the main functions of taxation—notably, raising revenue, redistributing income, and correcting externalities—and the mapping between these functions and various forms of taxation. In presenting a unified view, one grounded directly in a standard social welfare function, it should help expose and clarify connections among particular subjects in ways that often are beyond the purview of more focused treatments that consider, in much greater depth, a single piece of the larger puzzle.

Implicit in a conceptual approach is that empirical literature will not be a focus. Also excluded will be most aspects of tax incidence, questions of political economy, and macroeconomic issues. In other respects as well, this survey will not attempt to be comprehensive. Nevertheless, it covers a wide canvas and seeks to go into enough depth on the matters it does address to provide significant illumination.

Core features of the analysis appear in the preliminary sections. Section 2 considers the purposes of taxation, discusses the need for an integrated view that relates different policy instruments to specific objectives, and motivates and introduces the standard welfare economic approach to taxation. Section 3 presents optimal income taxation analysis, emphasizing the main conclusions, the intuitions underlying them, and the results of simulations.

Section 4 extends the analysis of section 3 to consider optimal commodity taxation in a setting in which an income tax is available. This extension proves particularly valuable in later sections because so many forms of taxation and other policies are analogous to differential commodity taxation. Some of the payoff appears in section 5, which considers other types of taxation, including income taxes that apply to capital as well as labor income (and the contrast between such income taxes and a personal consumption tax), corporate taxation, transfer (estate and gift) taxation, social security taxation, state and local taxation, and international taxation.

Because raising revenue and redistributing income are two central functions of taxation, a complete understanding requires further attention to government expenditures. Accordingly, section 6 analyzes income transfer payments and section 7 incorporates

[1] Many of the topics considered here are also examined in greater depth in *The Theory of Taxation and Public Economics* (Kaplow, in press), which has a similar motivation.

public goods into the framework. An additional function of taxation, the correction of externalities, is the subject of section 8.

A range of further topics are related to the central framework in the remaining sections. Section 9 examines a number of issues that arise in a dynamic setting: inflation, risk-bearing, transitions, capital gains, human capital, a lifetime horizon, and budget deficits and intergenerational redistribution. Section 10 addresses how different types of family units (single individuals, married couples, households with children) should optimally be treated relative to each other. Section 11 introduces problems of administration and enforcement. Section 12 briefly considers other important features of tax systems: the choice of the tax base and the differences among various forms of consumption taxation. Section 13 discusses tax equity, including the question whether social welfare functions should depend only on individuals' utilities, the choice of social welfare function, and other normative criteria sometimes suggested to be pertinent to tax policy.

2. Framework

2.1. *Purposes of taxation*

Raising revenue to fund government expenditures on public goods and services is a fundamental purpose of taxation. This task of raising revenue is intimately related to the second purpose of taxation, achieving an acceptable distribution of income. The reason is that, if all individuals were identical or if raising revenue was the only objective, the revenue need could be met in developed economies without distortion through the use of a uniform, lump-sum tax, sometimes referred to as a head tax or poll tax. Substantial reliance on constant per capita levies is unacceptable precisely because of distributive concerns.[2] And once distributive concerns are admitted, it is familiar that economic distortion becomes a central problem. Hence, using tax and other instruments to optimize the tradeoff between distribution and distortion is a principal focus of the economic analysis of taxation.

Taxation is also employed to achieve additional goals. The correction of externalities will be considered in section 8, while other objectives, notably economic stabilization, are beyond the scope of this survey.

2.2. *Integrated view*

To analyze a type of taxation or a particular tax reform proposal, it is helpful to bear in mind a number of considerations that involve the relationships among various components of the fiscal system. First, it is important to specify a policy completely rather

[2] See also subsection 4.4 on Ramsey taxation.

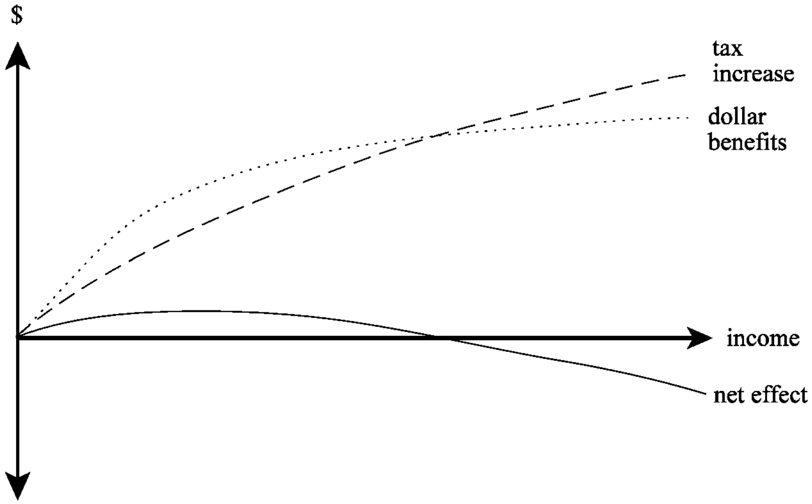

Figure 2.1. Government expenditure financed by gasoline tax increase.

than to consider individual pieces in a vacuum. For example, a gasoline tax increase may appear to be moderately regressive, which is to say that the average tax burden may increase less than proportionately with income.[3] See the dashed line, "tax increase," in Figure 2.1. A tax increase, however, generates revenue that upsets budget balance, so a complete specification of this policy requires identification of how the funds will be spent. Suppose that the revenue will be expended on a public good (or a reduction in some other tax), and that the incidence is favorable to the rich, but to an extent that is less than proportionate with income. See the dotted line, "dollar benefits," in Figure 2.1. The net effect, depicted by the solid line, is to redistribute toward poor and moderate-income individuals. Hence, what appeared to be a regressive gasoline tax increase, considered in isolation, has a net redistributive effect. Obviously, a different conclusion could be reached with different assumptions about expenditures, and the same point holds if the initial tax increase had instead been proportional or progressive. Because individual tax changes (and expenditure decisions) are part of a larger system with many instruments that may be adjusted in various ways, it is often unhelpful and potentially misleading to characterize any one instrument in a vacuum.

Second, particular policy instruments, such as forms of taxation, should be matched to those objectives to which they are most suited. For example, if consumption of gasoline

[3] Progressive, proportional, and regressive taxes are ordinarily defined as ones whose *average* rates rise, are constant, or fall with income. Occasionally, these terms are associated with *marginal* rates, but that usage will not be followed here. A motivation for focusing on average rates is that "progressive" taxes are often associated with redistributive taxes, and as subsection 3.2 will make clear, a tax with constant marginal rates (a flat tax) can be highly redistributive, in which case it will have rising average rates.

causes pollution, a gasoline tax would likely be a superior means of correcting this externality than an income tax, although the latter does tend to reduce consumption as a whole, including the consumption of gasoline. Conversely, if the objective is income redistribution, an income tax is likely to be more appropriate than a gasoline tax (which, as the preceding example indicates, is capable of income redistribution in combination with other instruments).

It turns out that most types of taxation are optimally utilized in specialized ways. A general income tax (or personal consumption tax) tends to be best to address redistribution, while most other forms of taxation are primarily justified because they target particular externalities or other imperfections, or because they address administrative and enforcement problems associated with other taxes. Although it is familiar that addressing a specific externality is best accomplished, if feasible, with a highly focused instrument, such as a corrective tax based on the externality itself, the notion that redistribution should be addressed almost exclusively with the income tax is less widely understood and thus deserves some further elaboration.

Consider, for example, whether luxury taxes should be employed to aid in the redistribution of income. (A complete analysis appears in section 4, on commodity taxation.) Initially, observe that any redistribution thereby accomplished could instead have been achieved with an adjustment to the income tax. That is, whatever is the incidence of the luxury taxes across the income distribution, one instead could have modified the income tax schedule to obtain the same result. Moreover, the use of luxury taxes tends to be a less efficient means of generating the same extent of redistribution. The reason is that luxury taxes distort both the consumption choices of the rich—who are induced to shift away from the taxed luxuries—and also the labor-leisure choice of the rich for, just as with an income tax increase, the effect of luxury taxes is to reduce the earner's benefit from additional labor effort. This lesson generalizes to other forms of taxation (and to government expenditures and regulation; see sections 7 and 8).

Given this conclusion, it is often useful to assess tax and other policy changes—other than pure reforms of the income tax and transfer system—using a *distribution-neutral* approach, as outlined in Kaplow (1996c, 2004, 2006b). That is, for any given policy, say a proposed increase in luxury taxes or in the gasoline tax, one can imagine that it is accompanied by an offsetting adjustment to the income tax and transfer system—one that, as a whole, keeps the distribution of utility constant. When such a policy experiment is examined, the relevant effects will tend to be solely the efficiency consequences regarding the specific target of the instrument in question: reduction in the consumption of luxuries or in the use of gasoline. In the former case, this consequence would tend to be inefficient (assuming the absence of externalities) whereas in the latter case the result would enhance efficiency (assuming that the externalities to gasoline consumption were not already fully internalized). The question in assessing the desirability of various forms of taxation then becomes, for individuals at a given level of income: Do we wish to relatively discourage—or in the case of subsidies or selective tax exemptions, relatively encourage—particular behaviors? For example, in examining the taxation of transfers (gift and estate taxes), questions of distribution, labor supply, and revenue-raising can largely be cast aside—for these are held constant by the offsetting income tax adjustment—and one would focus instead on whether it is desirable to discourage private income transfers relative to expenditures on direct consumption for oneself.

Relatedly, the foregoing distribution-neutral approach is extremely useful in examining policy packages that may not be distribution-neutral. In such cases, one can perform the following two-step decomposition: (1) Implement the target policy with a hypothetical adjustment to the income tax and transfer schedule that is distribution-neutral overall. (2) Implement a further reform that replaces the foregoing income tax adjustment with the one in the originally specified (and non-distribution-neutral) policy package. Step 1 can be analyzed as suggested previously. Step 2, it should be observed, is a purely redistributive adjustment to the income tax. Accordingly, for a vast range of policy packages—involving various mixes of taxation, expenditure, and government regulation—one can employ a generic approach to step 2. Furthermore, the necessary analysis for this step is the same as that required to assess pure questions of redistribution, as developed in section 3.

Use of a distribution-neutral approach (employing, where necessary, the proposed two-step decomposition) has many virtues. Most important, it greatly facilitates the analysis of the intrinsic effects of a policy, permitting specialization by analysts and comprehension of results by policy-makers. Note that if this approach is eschewed, anyone analyzing a gasoline tax increase, for example, would not only have to determine and assess the intrinsic effects of taxing gasoline, but would also have to determine what degree of redistribution should be assumed to accompany the reform, undertake an analysis of this redistribution (including the choice of labor supply elasticities and other parameters), and choose a social welfare function (SWF) to evaluate the consequences.[4] Likewise, two studies of a given gasoline tax increase could reach different conclusions for a variety of reasons that could prove difficult to untangle. Indeed, different conclusions are likely even if the studies agree on the intrinsic effects of the gasoline tax increase, that is, on the analysis of step 1 of the decomposition. A further benefit of this separation is for policy-makers, who may well wish to make their own

[4] It is common for analysts of other policy reforms to choose an income tax adjustment in a simple but essentially arbitrary manner, for example, by assuming that individuals' tax burdens adjust by a constant amount or proportionately. There is no accepted standard approach, and those most commonly used often involve redistribution (whether more or less redistribution depends on the target policy under consideration). Analysts also may consider actual reform proposals, although these often evolve and themselves may be incomplete (for example, they may not involve budget balance but instead increase a deficit that must in principle be financed by future tax adjustments). Although this survey does not consider matters of political economy, it should be noted that the political assumption implicit in the distribution-neutral approach—that the particular reform in question will not change the existing equilibrium of political forces with regard to the extent of redistribution—appears more plausible (on average and over time) than an arbitrary specification of how redistribution would change or an assumption that all reforms, regardless of their individual or cumulative distributive effects, would be financed in a particular, pre-specified manner.

choices regarding redistribution, applying their own assessments thereof. The two elements of the decomposition can in fact be implemented independently of each other, and enactment of only a single component will often be sensible, notably, if the intrinsic policy is efficient but the redistributive effect is deemed undesirable, or vice versa. For all of these reasons—although primarily for the greater conceptual clarity that results—a distribution-neutral approach will be utilized in much of this survey (but obviously not when analyzing pure redistribution) in attempting to illuminate the distinctive features of various forms of taxation.

2.3. Social objective

Evaluation of purely redistributive changes to the tax system, the focus of section 3 on optimal income taxation, requires specification of the social objective, in the guise of a social welfare function. The need for an explicit statement of the social objective is heightened by a number of considerations: Not all reforms affecting distribution can readily be classified as more or less redistributive (replacing a graduated income tax with a flat tax may benefit both the poor and rich at the expense of the middle class), subtle effects on distribution are caused by important tax policy choices (adjusting the accuracy of the tax system will increase the tax burdens of some and reduce those of others), and heterogeneity (especially among different types of family units) is an important feature bearing on redistribution in complex ways.

Despite the need for explicit use of a social welfare function, tax policy analysis has often adopted a looser approach. Standard treatments such as Musgrave and Musgrave (1973) and Stiglitz (2000) list multiple objectives of tax policy, like efficiency, fairness or equity (itself consisting of various dimensions or principles), revenue adequacy, simplicity, and administrability. Some of these criteria seem to be proxies for or subsets of others (simplicity is not a good in itself, but bears on efficiency and fairness) and others, especially various notions of fairness (such as "ability to pay") are notoriously vague, subject to competing interpretations, and in some instances largely free of content.

Mirrlees's (1971) seminal contribution on optimal income taxation, it should be noted, was motivated in significant part by the desire to link positive analysis of the effects of taxation to a normative framework that allowed for a rigorous synthesis of concerns for efficiency and distribution. This framework is provided by the standard welfare economic approach of basing all policy assessment on effects on individuals' utility and employing a social welfare function to aggregate individuals' utilities to make a comprehensive appraisal. This approach will be outlined here and followed throughout this essay. The justification for focusing exclusively on individuals' well-being, the choice of social welfare function, and the possible relevance of other equity criteria will be considered in section 13.

A social welfare function $SW(x)$ indicates how any regime or social state x (taken as a complete description thereof) is evaluated, where higher values indicate superior outcomes. Here, we are concerned with so-called individualistic SWFs, wherein social welfare depends only on individuals' utility or well-being. The functional form of

SW incorporates a view of distributive justice. In the present context of assessing redistributive taxation, it is standard to use an additive form that assumes a continuous population.

$$SW(x) = \int W\big(u_i(x)\big) f(i)\, di, \tag{2.1}$$

where u is a utility function, the subscript i indexes individuals' types, and $f(i)$ is the density of type i individuals in the population. The functional form of W on the right side of (2.1) incorporates a view of distributive justice, as can be seen from the following common formulation.

$$\begin{aligned} SW(x) &= \int \frac{u_i(x)^{1-e}}{1-e} f(i)\, di, \quad \text{for } e \neq 1 \\ &= \int \ln u_i(x) f(i)\, di, \quad \text{for } e = 1, \end{aligned} \tag{2.2}$$

where e indicates the degree of aversion to inequality in the distribution of utility levels.[5] If $e = 0$, social welfare is the sum (integral) of utilities, so the SWF is utilitarian. Higher levels of e correspond to increasing degrees of social aversion to inequality in the distribution of utilities. In the limiting case, as e approaches infinity, one has the maximin formulation associated with Rawls (1971) under which all weight is placed on the utility of the least-well-off individual.

It is useful to distinguish between two sources of aversion to inequality in the distribution of incomes. First, there is concavity in individuals' utilities as a function of consumption. To focus on this feature, consider the utilitarian SWF ($e = 0$). Furthermore, consider the case in which (abstracting from the effect of labor effort on utility) individuals' utility functions are given by $\ln c$, where c denotes consumption. Marginal utility equals $1/c$, so the marginal utility of a poor person with consumption of \$10,000 is ten times that of an upper-middle-income person with \$100,000 and one hundred times that of a rich person with consumption of \$1,000,000. If one considered a utility function with constant relative risk aversion of 2 (instead of 1, as in the preceding case), marginal utility would equal $1/c^2$; then these multipliers would be one hundred and ten thousand respectively. These factors indicate how much distortion would be tolerable in redistributing income: For example, when the factor is ten, further redistribution would raise social welfare as long as less than 90% of what the higher-income individual pays is lost in the redistributive process. Clearly, concavity of individuals' utility functions is an important source of a social preference for redistribution.[6]

[5] To motivate the latter version in (2.2), for the case in which $e = 1$, the numerator in the former may alternatively be written as $u_i(x)^{1-e} - 1$ (subtracting the constant having no effect on the ordering of states). Then, taking the limit as e approaches 1 (using l'Hôpital's rule) yields the latter expression.

[6] As is familiar from Edgeworth (1897), any concavity in u would, but for incentive and any other cost concerns, be sufficient to warrant complete equalization in individuals' levels of consumption.

Second, concavity in the SWF itself—in the W function in (2.1), corresponding to $e > 0$ in (2.2)—further favors redistribution. The relative importance of this factor will depend on the concavity of individuals' utility functions. If they are highly concave, then concavity in W may not contribute that much more to the social preference for equality.

Analysts sometimes, such as in performing optimal income tax simulations, use a single concavity parameter to refer to the overall concavity of social welfare as a function of individuals' consumption, in which case one may interpret any results as produced by varying combinations of concavity in the underlying u and W functions. Nevertheless, the two sources of concavity are conceptually distinct: The degree of concavity in u is an empirical question, whereas the degree of concavity in W is a normative matter.

For most of this essay, the degree of concavity in either u or W will not have a qualitative effect on the analysis. In section 3, addressing the optimal extent of redistribution, concavity will obviously be quantitatively important. In most other sections, there will not even be a quantitative effect because, as subsection 2.2 explained, the extent of redistribution will be held constant. However, in addressing some topics, such as in section 10 on taxation of different family units, it turns out that the extent of concavity may have qualitative effects, for subtle reasons that will be elaborated.

3. Optimal income taxation

3.1. Model

The analysis of optimal income taxation addresses the question of how an income tax should be designed in order to maximize a standard SWF subject to a revenue constraint, thus integrating consideration of the revenue-raising and distributive objectives of taxation. The standard model considers a one-period setting in which individuals' only choice variable is their degree of labor effort, there is a single composite consumption good, and government expenditures on public goods are taken as given. A variety of extensions will be examined in subsequent sections.

An individual's utility is given by $u(c, l)$, where c denotes consumption, l denotes labor effort, $u_c > 0$, and $u_l < 0$.[7] An individual's consumption is given by

$$c = wl - T(wl), \tag{3.1}$$

where w is the individual's wage rate and T is the tax-transfer function (usually referred to simply as a tax function or schedule). Each of these components deserves further elaboration.

[7] Much literature on optimal labor income taxation expresses utility as a function of leisure, or $1 - l$, where "1" denotes a normalized available amount of time for each individual. Additionally, it is common to use indirect utility functions, perhaps expressed as a function of lump-sum or virtual income and of a net-of-tax wage rate. Although these devices offer advantages, for purposes of the present exposition the use of direct utility expressed as a function of consumption and labor minimizes notation and is more transparent.

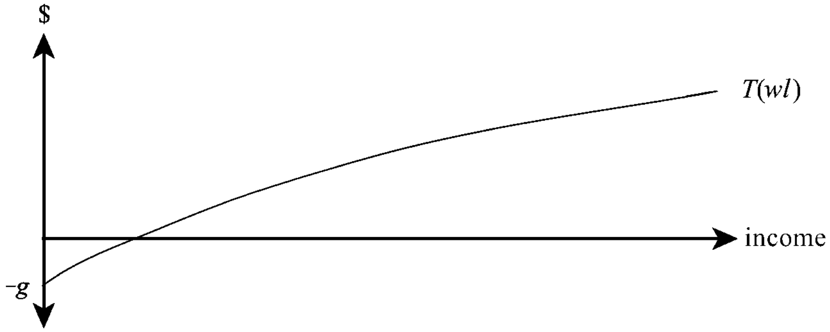

Figure 3.1. Nonlinear income tax and transfer schedule.

The motivation for redistributive taxation is that individuals differ, in particular in their wages, that is, their earning abilities. The distribution of abilities will be denoted $F(w)$, with density $f(w)$, the population being normalized to have a total mass of one. Individuals' abilities are indicated by their given wage rate, taken to be exogenous. Their pre-tax earnings are the product of their wage rate and effort level. More broadly, one can interpret effort as including not only hours of work but also intensity, and not only productive effort but also investments in human capital.

Taxes and transfers, $T(wl)$, at any income level may be positive or negative. The (uniform) level of the transfer received by an individual earning no income, that is, $-T(0)$, is usually referred to as the grant g. See Figure 3.1.

The tax schedule $T(wl)$ is taken to represent the entire tax-transfer system. Taxes may include sales taxes or value-added tax (VAT) payments in addition to income taxes. Transfers include those through the tax system, such as the Earned Income Tax Credit (EITC) in the United States, welfare programs (see section 6), and under some interpretations public goods (see section 7).[8]

Taxes and transfers are taken to be a function of individuals' incomes, assumed to be observable, and it is this dependence of taxes on income that is the source of distortion. If taxes could instead depend directly on individuals' abilities, w, individualized lump-sum taxes would be feasible and redistribution could be accomplished without distorting labor supply. Ability, however, is assumed to be unobservable.

Individuals choose the levels of labor effort l that maximize $u(c, l)$ subject to their budget constraints (3.1). An individual's first-order condition is

$$w\big(1 - T'(wl)\big)u_c + u_l = 0, \tag{3.2}$$

where a prime denotes the derivative with respect to a function's only argument. In this case, $T'(wl)$ indicates the marginal tax rate of an individual earning income of wl.

[8] The inclusion of transfers is extremely important both practically, since they are in fact significant, and conceptually, since otherwise redistribution would be limited to transfers between the rich and middle class, once the poor were exempted from the tax system.

The government's problem is taken to be the choice of a tax-transfer schedule $T(wl)$ to maximize social welfare, which (appropriately modifying expression (2.1)) can be expressed as

$$\int W\bigl(u\bigl(c(w), l(w)\bigr)\bigr) f(w)\, dw, \tag{3.3}$$

where c and l are each expressed as functions of w to refer to the level of consumption achieved and labor effort chosen by an individual of type (ability) w. This maximization is subject to a revenue constraint and to constraints regarding individuals' behavior. The former is

$$\int T\bigl(wl(w)\bigr) f(w)\, dw = R, \tag{3.4}$$

where R is an exogenously given revenue requirement.[9] Here, revenue is to be interpreted as expenditures on public goods that should be understood as implicit in individuals' utility functions; because these expenditures are taken here to be fixed, they need not be modeled explicitly. Regarding the latter constraints, individuals are assumed to respond to the given tax schedule optimally, as described by their first-order conditions (3.2), which determine the functions $c(w)$ and $l(w)$.[10]

Mirrlees's (1971) original exposition has been followed by subsequent elaborations, much of which is synthesized in Atkinson and Stiglitz (1980), Stiglitz (1987), Tuomala (1990), and Salanié (2003). Because the problem is formidable, the present survey will be confined to stating basic results, such as are embodied in first-order conditions and produced by simulations.

3.2. Linear income tax

A linear income tax is defined as a tax schedule

$$T(wl) = twl - g, \tag{3.5}$$

where t is the (constant, income-independent) marginal tax rate and g, as previously noted, is the uniform per-capita grant. For example, consider the linear (flat) tax depicted in Figure 3.2.

[9] Some of the literature equivalently expresses this constraint in terms of aggregate resource balance, which requires that the sum of resources devoted to private and public goods equals the amount produced by all individuals' labor efforts.

[10] Substituting individuals' first-order conditions can be problematic when there may be multiple local optima, as recognized and addressed by Mirrlees (1971). Although much subsequent work sets aside such complications, the matter is potentially important because, as will be seen, optimal tax schedules can involve falling marginal tax rates, which produce nonconvexities. In such instances, changing marginal rates can cause individuals to "jump" to a different level of income, a phenomenon found to be important in Slemrod et al. (1994).

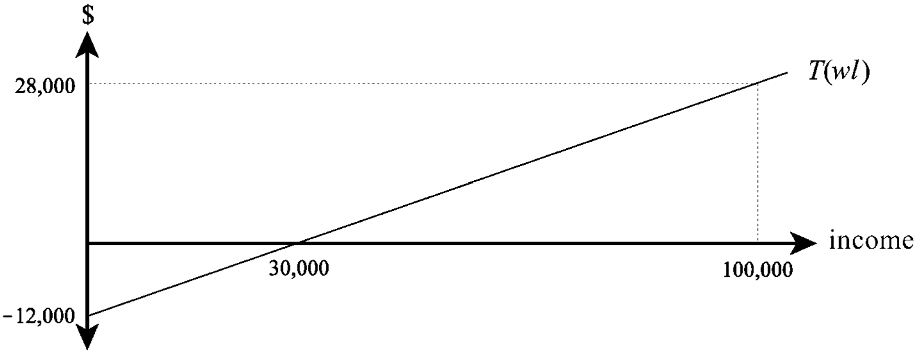

Figure 3.2. Linear income tax schedule, $t = 40\%$ and $g = \$12{,}000$.

In the literature, the schedule $T(wl)$, as mentioned, refers to a unified tax-transfer schedule. Note that this can be reinterpreted to align more closely with existing institutions and understandings. For example, the portion of the schedule to the right of \$30,000 of income can be understood as an ordinary (positive) flat or proportional income tax, with a marginal rate of 40% and an exemption for the first \$30,000 of income. The portion to the left of \$30,000 can be viewed as a transfer program having a value of \$12,000, a 40% phase-out rate, and a breakeven point of \$30,000. (Numerous other interpretations are also possible, including transfers that are not fully phased out until after \$30,000 but with an income tax exemption of less than \$30,000.) Further elaboration regarding transfers will be offered in section 6.

Expression (3.5) and Figure 3.2 also help illustrate how the degree of redistributiveness is not intimately connected to whether an income tax has graduated rates. Suppose, for example, that $t = 0$ and $g = 0$ (and that there is no revenue requirement). The result would be a totally nonredistributive flat tax: $T(wl)$ would be a horizontal line coincident with the x-axis. Now suppose that $t = 100\%$ and g is set equal to mean income (ignoring incentive effects). This would be a completely redistributive flat tax: $T(wl)$ would be a 45-degree line intersecting the x-axis at mean income and the y-axis at negative of the mean income. Hence, a purely proportional tax covers the full range of redistributive possibilities. It follows that nonlinearities in an optimal tax schedule, considered in subsections 3.3 and 3.4, will have less to do with the extent of redistribution and more to do with accomplishing redistribution in a more efficient manner (although the two dimensions are obviously interrelated).

To derive the optimal linear income tax, the government's maximization problem can be written in Lagrangian form as choosing t and g to maximize

$$\int \bigl[W\bigl(u\bigl((1-t)wl(w)+g, l(w)\bigr)\bigr) + \lambda\bigl(twl(w) - g - R\bigr)\bigr] f(w)\,dw, \qquad (3.6)$$

where λ is the shadow price of revenue, referring to the constraint (3.4), and (3.5) is substituted into (3.1) so that consumption is expressed in terms of the specific linear tax system under consideration. The first-order condition for the optimal tax rate can usefully be expressed as

$$\frac{t}{1-t} = -\frac{\operatorname{cov}(\alpha(w), y(w))}{\int y(w)\varepsilon(w) f(w)\, dw}, \tag{3.7}$$

where $y(w) = wl(w)$, income earned by individuals of ability w; $\varepsilon(w)$ is the compensated elasticity of labor effort of individuals of ability w; and $\alpha(w)$ is the net social marginal valuation of income, evaluated in dollars, of individuals of ability w.[11] Specifically with regard to the latter,

$$\alpha(w) = \frac{W' u_c(w)}{\lambda} + tw\left(\frac{\partial l(w)}{\partial g}\right). \tag{3.8}$$

The numerator of the first term on the right side of (3.8) indicates how much additional (lump-sum) income to an individual of ability w contributes to social welfare—u_c indicates how much utility rises per dollar and W' indicates the extent to which social welfare increases per unit of utility—and this is converted to a dollar value by dividing by the shadow price of government revenue. The second term takes into account the income effect, namely that giving additional lump-sum income to an individual of ability w will reduce labor effort ($\partial l(w)/\partial g < 0$), which in turn reduces government tax collections by tw per unit reduction in $l(w)$.

Expression (3.7) indicates how various factors affect the optimal level of a linear income tax. Beginning with the numerator, a higher (in magnitude) covariance between α and y favors a higher tax rate. In the present setting, $\alpha(w)$ will (under assumptions ordinarily postulated) be falling with income. Note that a larger covariance does not involve a closer (negative) correlation but rather a higher dispersion (standard deviation) of α and y. The dispersion of α will tend to be greater the more concave (egalitarian) is the welfare function W and the more concave is utility as a function of consumption (i.e., the greater the rate at which marginal utility falls with income). Income, y, will have a higher dispersion (again, under standard assumptions) when the distribution of underlying abilities is more unequal. In sum, more egalitarian social preferences, greater individual aversion to risk (more rapidly declining marginal utility of consumption), and higher underlying inequality will all contribute to a higher optimal tax rate.

The denominator of (3.7) indicates that a higher compensated labor supply elasticity favors a lower tax rate. The other terms in the integrand indicate that, ceteris paribus, the labor supply elasticity matters more with regard to high-income individuals and at ability levels where there are more individuals (typically the middle of the income

[11] There are many derivations of this condition, and it is expressed in a variety of equivalent ways. The present notation and manner of expression is close to that in Stiglitz (1987), page 1016, expression (29), and his derivation appears in note 31. See also Atkinson and Stiglitz (1980, pp. 407–408). These derivations, it should be noted, typically do not take into account that some individuals (those of low ability) will choose not to work, in which case (3.2) no longer characterizes their behavior (because they are at a corner solution). This problem is more often addressed in analyses of the optimal nonlinear income tax and in simulations.

distribution) because of the greater sacrifice in revenue. Note further that, if this compensated elasticity is taken to be constant, as is common in performing simulations, then the denominator is just the elasticity weighted by average income.

The foregoing exposition is incomplete in not emphasizing the various respects in which income effects are relevant (they influence α and also λ) and in ignoring that the values on the right side of (3.7) are endogenous. Especially for the latter reason, the literature has relied heavily on simulations.

The most-reported optimal linear income taxation simulations are those of Stern (1976). For his preferred case—an elasticity of substitution of 0.4,[12] a government revenue requirement of 20% of national income, and a social marginal valuation of income that decreases roughly with the square of income—he finds that the optimal tax rate is 54% and that individuals' lump-sum grant equals 34% of average income. (To put these figures in perspective, it should be understood that these estimates refer to the combination of all taxes; all government expenditures and all redistribution are financed by this single tax.) To illustrate the benefits of redistribution, he finds that a scheme that uses a lower tax, just high enough to finance government programs (that is, with a grant of zero), produces a level of social welfare that is lower by an amount equivalent to approximately 5% of national income.

Stern considers a number of other variations. If there is virtually no weight on equality, the optimal tax rate is only 25%, whereas if there is extreme weight on equality, specifically, the maximin case, the optimal tax rate is 87%. Returning to his central case, an extremely low labor supply elasticity implies an optimal tax rate of 79%, and an elasticity as high as had been used in some earlier literature implies an optimal tax rate of 35%. Additionally, his central estimate assumes that (nonredistributive) government expenditures are approximately 20% of national income. In the absence of the need to finance such expenditures, the optimal tax rate is 48%, and if expenditures were twice as high, the optimal tax rate is 60%.

3.3. Two-bracket income tax

Before proceeding to the general optimal nonlinear income tax problem, it is illuminating to consider briefly a simpler extension. A two-bracket income tax applies a constant rate t_1 to all income up to some specified level y° and another constant rate t_2 to all income over the specified level y°. See Figure 3.3 for an illustration in which $t_1 > t_2$. Here, the government chooses t_1, t_2, y°, and g to maximize social welfare.

This problem has been explored by Slemrod et al. (1994). They report simulations for an optimal two-bracket income tax using functional forms and parameters similar to

[12] In many simulations, including this one by Stern, investigators calibrate labor supply responsiveness by the elasticity of substitution between consumption and labor in a CES (constant elasticity of substitution) utility function. Such elasticities do not directly correspond to a compensated or uncompensated elasticity of labor supply. In fact, Stern's 0.4 elasticity of substitution corresponds to a case in which the uncompensated labor supply elasticity is negative.

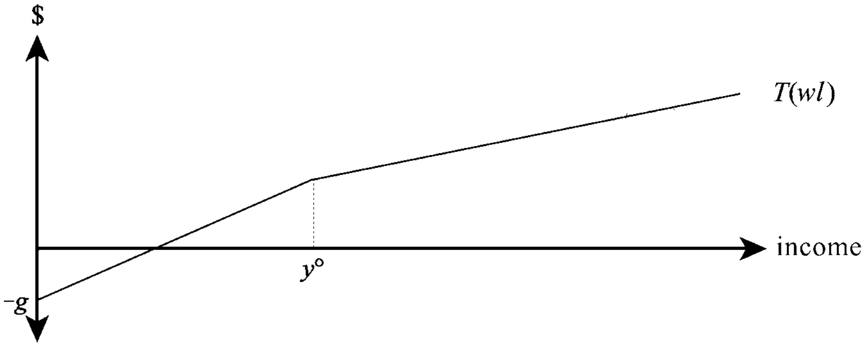

Figure 3.3. Two-bracket income tax schedule.

those employed by Stern (1976) and others. In all of the cases they consider, the optimal upper-bracket marginal tax rate is less than the optimal lower-bracket rate. Nevertheless, in all simulations in which the optimal transfer, g, is positive, the overall income tax schedule is progressive, which one should recall is defined as exhibiting rising average tax rates. In the case closest to Stern's central case, the optimal linear income tax has a rate of 58% whereas the optimal two-bracket tax has a marginal rate of 60% on low incomes and 52% on high incomes.

The intuition behind their results is that the lower rate on high-income individuals induces greater labor effort and thus raises more revenue without having to sacrifice revenue on income subject to the lower-bracket rate. This allows a larger grant g to be financed. Put another way, raising the bottom rate by Δt_1, while keeping the top rate fixed, is inframarginal regarding upper-bracket individuals; it collects $\Delta t_1 y^\circ$ from them without distorting their labor supply. Indeed, there is also an income effect on upper-bracket individuals that further increases their labor supply and thus revenue. Interestingly, as the social preference for equality increases, not only do the tax rates and level of grant increase, but the absolute size of the gap between the two tax rates widens in their simulations; that is, a greater preference for equality makes it optimal for the marginal rate on low-income individuals to be further above the marginal rate on high-income individuals. The intuition is essentially that just noted: Allowing the first rate to be higher enables additional revenue to be raised from high-income individuals to fund a higher transfer g, and this increase in g is relatively more valuable the greater the social benefit from redistribution.

3.4. Nonlinear income tax

Returning to the more general formulation of the optimal income taxation problem described in subsection 3.1 and depicted in Figure 3.1, the government chooses a tax schedule $T(wl)$ to maximize the SWF (3.3) subject to a revenue constraint (3.4) and constraints (3.2) requiring that individuals of all ability levels be maximizing their utility, taking the tax schedule as given. Mirrlees (1971) and subsequent investigators employ control-theoretic techniques to address this problem. In this maximization, the constraints regarding individuals' maximizing behavior entail that no individual of any type w will prefer the choice specified for any other type w°. (Readers may recognize this problem as related to the revelation principle used in work on mechanism design.[13])

This analysis can be summarized in a first-order condition for the optimal marginal tax rate at any income level y^*, where w^* and l^* correspond to the ability level and degree of labor effort supplied by the type of individual who would earn y^*. Following the presentation in Atkinson and Stiglitz (1980), who make the simplifying assumption that utility is separable between consumption and labor effort, adding the further assumption (discussed below) that marginal utility u_c is constant, conforming the notation, and engaging in some additional reshuffling, the condition can be expressed as[14]

$$\frac{T'(w^*l^*)}{1-T'(w^*l^*)} = \frac{1-F(w^*)}{\xi^* w^* f(w^*)} \frac{\int_{w^*}^{\infty}(1-\frac{W'(u(w))u_c}{\lambda})f(w)\,dw}{1-F(w^*)}, \tag{3.9}$$

where $\xi^* = 1/(1 + l^* u_{ll}/u_l)$—which, when marginal utility is constant as assumed here, equals $\varepsilon/(1+\varepsilon)$, where ε is the elasticity of labor supply. (This ε is often stated to be the compensated elasticity, but with constant marginal utility of consumption there is no income effect, so the compensated and uncompensated elasticities are identical.)

To aid in understanding expression (3.9), it is helpful to have in mind a simple perturbation of the income tax schedule that is used, for example, by Saez (2001). If one begins with some tax schedule $T(wl)$, assumed to be optimal, it must be that no slight adjustment to the schedule will change the level of social welfare. Consider an adjustment that slightly raises the marginal tax rate at some income level, y^* (say, in a small interval from y^* to $y^* + \delta$), leaving all other marginal tax rates unaltered. There are two

[13] Relatedly, following Stiglitz (1982a), many have advanced intuition and derived results by considering models with a finite number of types of individuals, often two. (This analysis parallels similar work on adverse selection in insurance models and on nonlinear pricing.) Corresponding incentive-compatibility constraints require that individuals will not wish to mimic other types, the problem in the case of redistributive taxation usually being that high-ability types may wish to mimic low-ability types in order to pay lower taxes.

[14] The relationship between Atkinson and Stiglitz's (1980) expression (13-54) on page 417 and that in the text is entirely straightforward except that their term ξ^* appears in the numerator rather than in the denominator. The difference in how ξ^* is defined (that here is the reciprocal of theirs) accounts for the difference in placement. The reason for the deviation is that it is convenient to follow convention and employ an ξ^* that corresponds more directly (and in particular is positively related) to the elasticity of labor supply. (Additionally, the assumption that u_c is constant allows some further simplification.) Expression (3.9) and Atkinson and Stiglitz (1980) are essentially identical to Stiglitz (1987) (expression (25) on page 1007 and the expression in note 17 on page 1008), Diamond (1998) (expression (10) on page 86), Dahan and Strawczynski (2000) (expression (2) on page 682), and Auerbach and Hines (2002) (expressions (4.12) and (4.15) on pages 1381–82). It is also similar to the two formulations in Saez (2001, p. 215).

effects of such a change. First, individuals at that income level face a higher marginal rate, which will distort their labor effort, a cost. Second, all individuals above income level y^* will pay more tax, but these individuals face no new marginal distortion. That is, the higher marginal rate at y^* is inframarginal for them. Since those thus giving up income are an above-average slice of the population (it is the part of the population with income above y^*), there tends to be a redistributive gain.

Expression (3.9) can readily be interpreted in terms of this perturbation.[15] Begin with the first term. Revenue is collected from all individuals with incomes above y^*, which is to say all ability types above w^*; hence the $1 - F(w^*)$ in the numerator. This factor favors marginal tax rates that fall with income: As there are fewer individuals who face the inframarginal tax, the core benefit of higher marginal rates falls. In the extreme, if there is a highest known type in the income distribution, the optimal marginal rate at the top would be zero because $1 - F$ would be zero: A higher rate collects no revenue but distorts the behavior of the top individual.[16] However, when there is no highest type, known with certainty in advance, this result is inapplicable. Furthermore, even with a known highest type, simulations suggest that zero is not a good approximation of the optimal marginal tax rate even quite close to the top of the income distribution, so the zero-rate-at-the-top result is of little practical importance.[17]

Raising the marginal rate at a particular point distorts only the behavior of the marginal type, which explains the $f(w^*)$ in the denominator of the first term. For standard distributions, this factor is rising initially and then falling, which favors falling marginal rates at the bottom of the income distribution and rising rates at the top. The denominator also contains weights of ξ^*, indicating the extent of the distortion, and w^*, indicating how much production is lost per unit of reduction in labor effort. The elasticity is often taken to be constant, although some empirical evidence on the elasticity of taxable income (see subsection 11.3) supports a rising elasticity due to the greater ability of higher-income individuals to avoid taxes.[18] This consideration may favor marginal rates that fall with income. Finally, w^* is rising, which also favors falling marginal rates: The greater the wage (ability level), the greater the revenue loss from a given decline in labor effort.

The second term applies a social weighting to the revenue that is collected. The integrand in the numerator is the difference between the marginal dollar that is raised and

[15] Just as when interpreting the first-order condition (3.7) for the linear income tax, income effects and the endogeneity of terms being interpreted will be ignored. The latter problem is more serious here because parameters on the right side of (3.9) depend implicitly on marginal tax rates other than at y^* (through the term $W'u_c/\lambda$).

[16] This result first appears in Phelps (1973) and Sadka (1976) and is explored in some detail by Seade (1977). In (3.9), $1 - F$ also appears in the denominator of the second term; however, the integral in the numerator of the second term also equals zero. As w^* approaches its maximum, the second term as a whole approaches 1 minus the welfare weight on the top individual whereas the first term approaches zero.

[17] See, for example, Tuomala (1990).

[18] See, for example, Alm and Wallace (2000), Auten and Carroll (1999), Gruber and Saez (2002), and Moffitt and Wilhelm (2000).

the dollar equivalent of the loss in welfare that occurs on account of individuals above w^* paying more tax. As in the interpretation of (3.7), u_c is the marginal utility of income to such individuals, W' indicates the impact of this change in utility on social welfare, and division by λ, the shadow price on the revenue constraint, converts this welfare measure into dollars.[19] This integral is divided by $1 - F(w^*)$, which makes the second term an average for the affected population.

This term tends to favor marginal rates that rise with income. The greater is w, the lower is W' (unless the welfare function is utilitarian, in which case this is constant) and the lower would be the marginal utility of income u_c (had we not abstracted from this effect in the assumptions); hence at higher w^*, the average value of the term subtracted in the integrand is smaller, making the entire term larger. Note further that if social welfare or utility is reasonably concave, $W'u_c$ will approach zero at high levels of income, at which point this term will be nearly constant in w^*. That is, the term favors rising marginal tax rates when income is low or moderate, but has little effect on the pattern of marginal tax rates near the top of the income distribution.[20]

Because of difficulties in determining the shape of the optimal income tax schedule by mere inspection of the first-order condition (3.9), analysts beginning with Mirrlees (1971) have used simulations to help join the theoretical analysis with empirical estimates of labor supply elasticities and of the distribution of skills or income in order to provide further illumination. The discussion here will emphasize how the shape of the optimal nonlinear income tax varies from linearity because subsection 3.2 on the optimal linear income tax already reports how the overall level of marginal tax rates is affected by various parameters of the problem. Tuomala (1990) offers a useful survey and set of calculations. Perhaps his most notable conclusion is that, in all the cases he reports, marginal tax rates fall as income increases, except at very low levels of income. Mirrlees's (1971) original calculations had displayed a similar tendency, but subsequent researchers questioned the extent to which this result may have depended on the social preferences he stipulated or the arguably high labor supply response he assumed. Subsequent work, however, suggests that a greater social preference for equality or a lower labor supply response tends to increase the level of optimal marginal tax rates but does not generally result in a substantially different shape.

Some more recent work explores further whether there exist circumstances in which optimal marginal tax rates rise with income. Kanbur and Tuomala (1994) find that when

[19] Another natural way to think of the experiment of raising the marginal rate $T'(w^*l^*)$ is to suppose further that the additional revenue will be used to increase the uniform grant. The marginal social value of increasing the grant will, at the optimum, necessarily equal the shadow price λ of government revenue.

[20] Brito and Oakland (1977) and Seade (1977) showed that the optimal marginal tax rate at the bottom of the distribution is zero, a phenomenon that can be understood by reference to this term: If the higher marginal rate applies to literally everyone, so they all pay the same increment in tax, then there is no redistribution, but there still is distortion of the lowest type, who is subject to a positive marginal rate. However, since it is typical that the optimum has all individuals below some low ability level not working, it is not in fact the case that there is no redistribution from applying a positive marginal rate to the lowest type who chooses to work, and Ebert (1992) shows that a positive marginal tax rate at the bottom is indeed optimal in this case.

inequality in individuals' abilities (wages) is significantly greater than previously assumed (but in ranges they suggest to be empirically plausible), optimal marginal tax rates do increase with income over a substantial range, although for upper-income individuals optimal marginal rates still fall with income. Diamond (1998) examines a Pareto distribution of skills, instead of the commonly used lognormal distribution, under which the $(1 - F)/f$ component of (3.9) rises more rapidly at the top of the distribution, and finds that optimal marginal tax rates are rising at the top. However, Dahan and Strawczynski's (2000) simulations indicate that Diamond's result was driven in large part by his additional assumption that preferences were quasi-linear, thus removing income effects. (Nevertheless, their diagrams do suggest that, consistent with Diamond's claim, moving from a lognormal to a Pareto distribution favors higher rates—still falling, but notably less rapidly—at the top of the income distribution.) Saez (2001), using income distribution data in the United States from 1992 and 1993, finds that the shape of the distribution of $(1 - F)/wf$ is such that optimal rates should fall substantially well into the middle of the income distribution, to an income of approximately $75,000, rise until approximately $200,000, and then be essentially flat thereafter.[21]

Another important conclusion in Mirrlees's (1971) original work is that the optimal nonlinear income tax is approximately linear. If this is true, it may be that there is little loss in social welfare if only a linear income tax (which may have administrative advantages) is used. Subsequent investigators report a range of cases in which the optimal nonlinear income tax departs more substantially from a linear tax, but they do not generally report how much welfare loss would be involved in using only a linear scheme.

An additional result from the simulations is that, at the optimum, a nontrivial fraction of the population does not work, and this fraction is larger when social preferences favor greater redistribution and when the labor supply elasticity is higher. This outcome should hardly be surprising because, as the analysis of (3.9) and the simulations suggest, high marginal rates tend to be optimal at the bottom of the income distribution, along with a sizable grant. Relatedly, little productivity and thus little tax revenue is sacrificed when those with very low abilities are induced not to work (whereas substantial revenue is raised from the rest of the population, for whom marginal tax rates on their first dollars of income are inframarginal).

[21] For example, in his simulation with a utilitarian welfare function, a compensated elasticity of labor supply of 0.5, and a functional form for utility that has income effects, his optimal schedule has a marginal rate near 80% at the bottom of the income distribution that falls to approximately 40% at $80,000, and then rises to nearly 70% at the upper end, where it roughly levels off. However, his functional form for utility has income effects that rise with income to such an extent that the uncompensated elasticity approaches zero as w increases, which favors higher marginal rates at the top than otherwise. See also Dahan and Strawczynski (2004) for further exploration.

3.5. Elaboration

3.5.1. Taxation of earning ability

The need to use a distortionary labor income tax to achieve distributive objectives is premised on the infeasibility of individualized lump-sum taxes based on individuals' earning ability, which would be nondistortionary. The assumption is that differences in earning ability are unobservable, so income, a signal of earning ability, is taxed instead. However, given that income taxation is distortionary and, as a result, society cannot fully meet its desired distributive objective, and that income itself is neither costlessly nor accurately observable (see section 11), it is worth considering the possibilities for basing taxation more directly on ability.

One strategy would be to attempt to observe individuals' wages or to infer wages from income and hours. Hours, however, are difficult to observe and both hours and wages are manipulable, such as by extending reported hours and lowering the reported wage (keeping earnings constant, and thus both employer and employee indifferent); self-employment poses a particularly serious problem. Another approach would be to measure proxies of earning ability, such as through testing. Unfortunately, skills measurable by testing explain only some of the variance in earnings ability. Furthermore, if taxes were to be based on test results or other ability measures, individuals would adjust their performance and thereby distort the measurement. A third technique—one sometimes employed—is to adjust taxes and transfers for observable attributes, such as physical disability, age, or family composition.

There has been little formal analysis of the taxation of earning ability. Stern (1982) compares an ability tax supplemented by a purely proportional income tax and an optimal nonlinear income tax. He assumes that there will be classification errors with an ability tax and considers how large the errors have to be to make the nonlinear income tax preferable. He finds that, the greater the preference for equality, the less attractive is an ability-tax scheme because mistakes in which low-ability individuals are misclassified as high types are more socially costly. Unfortunately, his comparison is not clean because he allows a more powerful (nonlinear) income tax when there is no ability tax; moreover, he uses a model with only two types of individuals, which further increases the relative power of a nonlinear income tax.

A broader approach would be to suppose that there exists an imperfect signal (or signals) of ability and allow the government to make the tax and transfer schedule a function of the signal. The signal, call it θ, could be an index of discrete classifications or a continuous variable. Then, the first-order condition for the optimal nonlinear income tax problem, expression (3.9), could be restated, showing that the tax schedule and the distribution and density functions also depend on θ, giving us $T(w^*l^*, \theta)$, $F(w^*, \theta)$, and $f(w^*, \theta)$, respectively.[22] (Note that the separate tax schedules would be

[22] One could also allow the utility function in expression (3.9) to depend on θ, recognizing, for example, that both utility levels and the marginal utility of consumption could be affected by such observable characteristics as disabilities or family composition, the latter case being the subject of section 10.

linked in a common optimization by the shadow price λ.) This approach will be used in subsection 6.2 to address the optimal form of categorical assistance under transfer programs. As a special case, if one supposes that one of the groups is homogeneous (all of one type), the optimal schedule for that group would involve a zero marginal tax rate, with all redistribution to or from the group accomplished through the group's lump-sum transfer. This is the structure of Akerlof (1978), in which he assumes that a subset of the lowest-ability group can be identified perfectly ("tagged").

3.5.2. Additional considerations

In addition to factors explored elsewhere in this survey, a number of considerations further complicate the optimal income taxation problem. One is that income may be a noisy signal of ability, whether because of variations in occupations (for a given ability, one job may pay more to compensate for specific disamenities) or in preferences (an individual may earn more not because of greater ability but rather due to a higher marginal utility of consumption or a lower marginal disutility of labor effort). Another possibility is that individuals may have preferences concerning redistribution itself, perhaps due to altruism or envy.[23] Other topics that have been explored include liquidity constraints[24] and general equilibrium effects of redistribution on the distribution of pre-tax wages.[25] Some of these factors may make redistribution more attractive than otherwise, some less attractive, and some are indeterminate without further specification of the model or parameter values. Most of these subjects have received only modest attention despite their potential importance to the optimal income taxation problem.

4. Commodity taxation

The analysis of labor income taxation may be extended by considering a setting in which consumption consists not of a single, composite good but a range of goods and services, and each type of consumption may be taxed or subsidized at its own rate.[26] Commodity

[23] See, for example, Hochman and Rodgers (1969) on the possibility that the rich benefit from redistribution to the poor, Pauly (1973) on redistribution as a local public good, Veblen (1899), Duesenberry (1949), Boskin and Sheshinski (1978), Frank (1984a, 1984b, 1985, 1999), and Tuomala (1990) on individuals' concern for status, and Easterlin (1973, 1974, 2001), Frank (1984b, 1985), and Veenhoven (1991) on the possibility that individuals' long-run preferences may be largely relative.

[24] See, for example, Hoff and Lyon (1995), Hubbard and Judd (1986) (and Hall's and Summers's comments thereon), and Polinsky (1974).

[25] See, for example, Feldstein (1973), Allen (1982), Carruth (1982), and Stiglitz (1982a).

[26] Commodity taxation was traditionally referred to as a form of indirect taxation, in contrast to an income tax (or personal consumption tax), which was described as direct taxation. The standard interpretation is that direct taxes can plausibly be tailored to individuals' circumstances, allowing notably for uniform per capita taxes or transfers and for nonlinear taxation. By contrast, indirect taxes, such as commodity taxes, are impersonal; they do not allow a uniform levy because individuals cannot (at least for purposes of indirect taxes)

taxation is important in its own right, for it is possible to tax commodities differentially and this is often done (e.g., taxes on gasoline, hotel stays, alcohol, and tobacco). Additionally, in general sales tax or value-added tax (VAT) systems, it is common to apply differential rates, such as by providing rate reductions or exemptions for purchases of food. Whether luxury taxes are desirable on redistributive grounds is another sort of question directly investigated in this section.

Moreover, commodity taxes are important conceptually because they provide a basis for analyzing (directly or by extension) a number of other subjects.[27] For example, taxation of savings can be viewed as differential taxation of future versus present commodities, and transfer (estate and gift) taxes are differential taxes on different forms of consumption by donors. Other subjects, including expenditures on public goods and corrective taxes, can also be analyzed by reference to the basic commodity taxation model. See subsections 7.2 and 8.3. As will become apparent, this section contains a formalization of the distribution-neutral approach presented in subsection 2.2 that is applicable to a broad range of tax and other governmental policies, including most of those that are not concerned exclusively with redistribution (i.e., the pure optimal income tax problem of section 3).

To foreshadow the results, the main conclusion for the basic case is that no differentiation in commodity taxes—equivalent to a system of no commodity taxes or subsidies—is optimal. This important result was established by Atkinson and Stiglitz (1976) for the case in which the nonlinear income tax is set optimally. The exposition here will follow Kaplow (2006b), who extends their result to the more general case in which one begins with an arbitrary nonlinear income tax in an intuitive manner that uses the previously described distribution-neutral approach.[28] The argument shows that, in a basic setting, if the income tax is adjusted to hold distribution constant, labor supply also remains unchanged, so the only effect of commodity taxation is on the allocative efficiency of individuals' consumption decisions. The optimal result, therefore, involves no differential taxation; indeed, the elimination of differential commodity taxation can be accomplished in a manner that results in a Pareto improvement. Likewise, any reform of a system of commodity taxes and subsidies in the direction of simple efficiency with regard to consumption choices can be implemented in a way that makes everyone better off.

be identified. Relatedly, nonlinear indirect taxation is presumed to be impossible because of the infeasibility of charging different rates that depend on the amount an individual consumes, which would require identification of who purchases commodities and also that resale (arbitrage between individuals whose different consumption choices lead them to face different marginal tax rates) be preventable.

[27] Perhaps the closest case involves tax preferences such as deductions, exemptions, or credits for particular activities (such as energy conservation) in income tax systems, which are similar to direct subsidies.

[28] This approach was first used by Hylland and Zeckhauser (1979) in arguing that distributive concerns should play no role in cost-benefit analysis and was developed further in Kaplow (1996c, 2004). See subsection 7.2. Other discussions of commodity taxation in the presence of nonlinear income taxation that may not be optimal include Konishi (1995) and Laroque (2005). Additionally, Atkinson and Stiglitz (1976) and Deaton (1979) characterize the restrictions on utility functions that are necessary for no differentiation to be optimal when the optimal income tax is restricted to be linear.

4.1. Model

The model employed in section 3 for studying income taxation can be modified to incorporate commodity taxation as well. Instead of a single, composite consumption good c, it is now supposed that individuals may spend their after-tax-and-transfer income, $wl - T(wl)$, on any of n commodities, $x_1, \ldots, x_n$. Commodity prices (which equal constant unit production costs measured in units of income and thus may be thought of as prices paid to competitive producers) for goods x_i are p_i and commodity taxes are τ_i (which may be subsidies, in which case they are negative). Individuals as consumers thus face net prices of $p_i + \tau_i$, assumed to be positive.

An individual's budget constraint, instead of that given in expression (3.1), is now

$$\sum (p_i + \tau_i) x_i(wl) = wl - T(wl), \tag{4.1}$$

where summations throughout are from i equals 1 to n and the notation $x_i(wl)$ denotes the level of x_i chosen by an individual of earning ability w (and l likewise implicitly refers to the labor effort of an individual of type w). The government's budget constraint, instead of (3.4), becomes

$$\int \left[T(wl) + \sum \tau_i x_i(wl) \right] f(w)\, dw = R. \tag{4.2}$$

Before undertaking the analysis, it is useful to discuss the relationship between the average overall levels of commodity taxation and of income taxation, and also related matters of normalization. Initially, observe that there are infinitely many equivalent ways to describe and implement any commodity tax system. To see this, consider uniform commodity taxes, that is, commodity tax schemes for which $\tau_i = \alpha p_i$, for all i. Compared to a baseline with no commodity taxation, if $\alpha > 0$, everyone pays proportionally more for any bundle of commodities. Such a commodity tax system is equivalent to the imposition of a linear income tax (or to a uniform adjustment of a preexisting, possibly nonlinear, income tax). To see this, examine the budget constraint (4.1) for this commodity tax system when there is no income tax at the outset.

$$\sum (p_i + \alpha p_i) x_i(wl) = wl. \tag{4.3}$$

Factoring $1 + \alpha$ outside the summation on the left, dividing both sides by $1 + \alpha$, and letting $t = \alpha/(1 + \alpha)$ yields

$$\sum p_i x_i(wl) = \frac{1}{1 + \alpha} wl = (1 - t) wl. \tag{4.4}$$

The left side of expression (4.4) is the cost of consumption in a world with no commodity taxes, and the right side is disposable income for the case of a linear income tax (with no grant). Introducing uniform commodity taxation is indeed equivalent to a uniform shift in the level of income taxation. Put in other words, a uniform consumption tax is equivalent to a linear tax on labor income, a simple result that is useful in examining the differences between consumption taxes and general income taxes (which also reach capital income) and also in understanding the relationships among various forms of consumption taxes, including a VAT. (See subsections 5.1.1 and 12.2.)

Because varying the overall level of commodity taxes is equivalent to varying the level of marginal tax rates under an income tax, which is the subject of the optimal income taxation literature surveyed in section 3, work on commodity taxes and subsidies has focused on the question of whether and when differential commodity tax rates are optimal.

4.2. Analysis

Assume that individuals' utility functions are weakly separable between labor (leisure) and all other commodities, taken together. That is, their utility functions can be expressed as $u(v(x_1, \ldots, x_n), l)$, where v is a subutility function. This formulation implies that, for a given level of after-income-tax income, individuals will allocate their disposable income among commodities in the same manner regardless of the level of labor effort required to earn that level of income. Put another way, the ratio of the marginal utilities of consumption for any two commodities, at given levels of consumption of those commodities and of all other commodities, is independent of the level of labor effort. (For commodities i and j, this ratio is simply $u_v v_i / u_v v_j = v_i / v_j$.) As will be seen, this further implies that changes in the allocation of after-tax income among commodities that are caused by commodity tax reforms (that are compensated in the sense of keeping utility constant) will not affect the choice of labor effort. This separability assumption will be discussed further in subsection 4.3.

Using this framework, a differentiated tax system $\{\tau_1, \ldots, \tau_n\}$, $T(wl)$ is one for which there exists i, j such that $(p_i + \tau_i)/(p_j + \tau_j) \neq p_i/p_j$. In other words, the ratio of net prices of at least one pair of goods does not equal its production cost ratio.

Assume that there exists some differential taxation and consider a commodity tax reform that eliminates all differentiation, specifically, by moving to a regime in which $\tau_i = 0$, for all i. Suppose that as an initial matter this commodity tax reform is combined with a distribution-neutral (offsetting) income tax adjustment that has the feature that every individuals' utility remains unchanged. Moving to the new commodity tax vector will tend to change individuals' utility because they no longer pay commodity taxes (or receive subsidies) and because, with a new relative price vector, they will change their consumption vectors. Whatever is the net effect on utility for any ability level w and given labor effort $l(w)$, define an intermediate income tax schedule $T^\circ(wl)$ at each income level so as to offset the net effect on utility. That is, examine an income tax schedule $T^\circ(wl)$ that has the property that, if all individuals (of every type w) continue to choose the same level of labor effort $l(w)$ as under the initial tax system, then their utility will be unchanged.[29]

[29] It is familiar to refer to this experiment as involving a (utility) compensated change, so at each level of income, wl, $T(wl) - T(wl)^\circ$ is the (Hicksian) compensating variation associated with the change in relative prices due to the commodity tax reform. (A difference is that, in the present formulation, labor supply is held constant, although it is to be demonstrated that this is indeed the case in any event.)

This reform, consisting of the elimination of commodity taxation and an offsetting income tax adjustment, can be shown to induce individuals to choose the same level of labor effort. Initially, observe that $T^\circ(wl)$ has the property that it leaves subutility v unaffected for all levels of income. That is, stated in reduced form, $V(wl) = V^\circ(wl)$ for all wl, where V is the maximized value of v (the value of v obtained at each wl when individuals choose the x_i's optimally, taking the commodity and income tax regime as given). This result about the subutility functions must be true because the income tax schedule $T^\circ(wl)$ is constructed such that $u(V(wl), l) = u^\circ(V^\circ(wl), l)$; because the function u does not change, the levels of subutility must be unchanged for each given level of l and thus of wl. To be sure, changing commodity taxes and changing the income tax schedule each will alter the level of subutility V produced by a given level of income wl; however, because of how the income tax schedule adjustment is constructed, these two sets of effects will be precisely offsetting. Furthermore, if the level of subutility V is unchanged for every possible level of income, then it also must be true that, for any choice of labor effort l, each type of individual's total level of utility is the same as it was before. In other words, $U(l(w)) = U^\circ(l(w))$ for all $l(w)$, where the reduced form $U(l(w))$ refers to the level of utility achieved for any choice of l by the given type w.[30] Since utility as a function of labor effort is precisely the same under the new, intermediate regime as it is under the initial regime, it follows that whatever level of labor effort $l(w)$ maximized $U(l(w))$ will also maximize $U^\circ(l(w))$. Accordingly, individuals will indeed choose the same level of labor effort under the newly constructed intermediate regime.

To complete the argument, consider the effect on revenue of the elimination of differential commodity taxation combined with the distribution-neutral tax adjustment involving the intermediate income tax schedule, $T^\circ(wl)$. The income tax adjustment, recall, derives from two effects of the commodity tax reform. First, the reform changes individuals' commodity tax payments, even assuming that they do not change their consumption decisions, and the income tax adjustment offsets this effect. Clearly, this combination will be revenue-neutral as a whole, because each type of individual's income tax payments rise or fall by just the amount that commodity tax payments fall or rise. Second, due to the changes in relative prices, individuals will be induced to change their consumption of various commodities. This change can only increase utility (for otherwise individuals would not choose to adjust their consumption choices). Hence, the income tax schedule adjustment that offsets this effect on utility will result in additional revenue being raised, generating a surplus.[31] Therefore, one can further adjust the income tax schedule to rebate this surplus, say in equal amounts to every individual. Because everyone's utility is the same under the intermediate regime and the initial regime,

[30] These functions, because denominated in utility, will differ among individuals with different earning abilities. However, with homogeneous preferences and weak leisure separability, as assumed here, the same tax adjustment (denominated in dollars) will work for all individuals.

[31] For a formal demonstration of the entire argument, see Kaplow (2006b).

it must be that, with the rebate, everyone's utility is greater in the resulting regime that eliminates differential commodity taxation than it is in the initial regime.

Accordingly, it is possible to eliminate differential commodity taxation in a manner that generates a Pareto improvement—specifically, by adjusting the income tax in a manner that produces a reform package that is overall distribution neutral. As shown in Kaplow (2006b), a similar approach can be used to show that proportional reductions in differential commodity taxation as well as other partial reforms that are efficient in the simple sense of reducing the amount of resources required to achieve individuals' initial levels of utility can be implemented in combination with an offsetting income tax adjustment so as to make everyone better off.[32] In the set of cases under consideration, a distribution-neutral reform also keeps labor supply constant; hence, the income redistribution problem, concerned with both distribution and labor supply distortion, can be separated from the commodity tax problem, which only affects individuals' choices among commodities. In other words, one can legitimately ignore distribution and labor supply because the reform packages under consideration hold both constant. When this is possible, it should not be surprising that standard efficiency principles indicate which commodity tax reforms are optimal.

4.3. Qualifications

There are a number of qualifications to the foregoing conclusion that differential commodity taxes and subsidies are inefficient. Most obvious is the case in which the consumption of some goods involves externalities, on which see section 8. Two additional qualifications concern the argument that a distribution-neutral (offsetting) income tax adjustment will not affect labor supply.

The assumption that individuals' utility is weakly separable in labor (leisure) rules out one source of possible labor supply effects. To illustrate the phenomenon, consider that taxing books, movie tickets, or swim suits (relative to other goods) tends to make leisure relatively less attractive. Given the distortion in favor of leisure caused by the income tax, this effect would be beneficial. Likewise, subsidizing substitutes for leisure, such as labor-saving devices, would tend to be advantageous. Because of the second-best setting that exists due to the assumed impracticality of taxing leisure directly (to offset the effect of taxing labor), it is optimal to distort other activities if (but only if) the distortion of the labor-leisure choice is thereby mitigated.[33] Some possible examples are identified in empirical work: Barnett (1979) finds that consumers substitute durable goods for leisure (and thus are a candidate for subsidies), Iorwerth and Whalley (2002) find that restaurant meals substitute for leisure whereas raw food complements leisure (implying that it may be optimal to reverse the common practice, superficially appealing

[32] Dixit (1975) and others characterize efficient partial reforms, although in a Ramsey model in which there is no concern for distribution and no income tax. See subsection 4.4.

[33] This point is first suggested by Corlett and Hague (1953), although in a Ramsey tax setting.

on distributive grounds, of taxing restaurant meals but exempting sales of raw food from sales taxes), and West and Williams (2007) find that gasoline is a leisure complement and thus should be taxed (by more than the level of the externalities associated with gasoline consumption).

As suggested by Mirrlees (1976), the inability to tax ability directly (see subsection 3.5.1) provides another possible basis for differential taxation in cases in which preferences for some commodities depend directly on individuals' abilities (rather than on their incomes, which reflect their abilities). Specifically, it tends to be optimal to impose a heavier burden on commodities preferred by the more able and a lighter burden on those preferred by the less able. For example, efficiency may favor taxing expenditures related to fine art (acquisitions of art objects, attendance at museums and the opera, and purchases of high-brow literature) and subsidizing simpler pleasures (bowling, attendance at professional wrestling, and viewing of trashy movies). Notice, however, that this argument does not imply that one should tax luxuries in general; because higher demand for luxuries is, by definition, a consequence of higher income, taxing luxuries distorts the labor-leisure decision. The present consideration is distinctive because it depends on preferences that vary with ability per se. Put somewhat differently, assuming two individuals were to earn the same income, the relevant question is whether the higher-ability person would, relative to the other, prefer a different mix of commodities.

The strong conclusion that reducing differential commodity taxation can be accomplished in a manner that yields a Pareto improvement is qualified by a number of additional considerations: heterogeneity of individuals' preferences, administrative and enforcement concerns, political economy considerations, and other factors. It should be emphasized, however, that most qualifications to the basic conclusion are largely orthogonal to standard redistributive considerations. Notably, one does not seek adjustments that are directly redistributive, such as by (relatively) taxing luxuries and subsidizing necessities. Indeed, as suggested by the foregoing example of expenditures on meals and on unprepared food, opposite adjustments may well be optimal.

4.4. *Ramsey taxation*

Most surveys and textbook treatments of optimal taxation devote substantial attention to Ramsey's model of taxation and the principles derived therefrom. See, for example, Atkinson and Stiglitz (1980), Auerbach (1985), Auerbach and Hines (2002), and Sandmo (1976). Ramsey's (1927) seminal paper addresses how to raise a given amount of revenue through commodity taxation when distributive considerations are ignored and an income tax is assumed to be unavailable. The familiar prescription is that, in the special case in which compensated demand schedules are independent (zero cross-elasticities), taxes should be inversely proportional to the elasticity of demand because distortion is less when the elasticity is lower.[34] The major qualification involves distribution, which favors higher taxes on goods consumed disproportionately by higher-income

[34] Another important result is that, with constant returns to scale in production or the availability of a 100% profits tax, production efficiency is optimal (i.e., no differential taxation of inputs). See Diamond and Mirrlees

individuals. See, for example, Atkinson and Stiglitz (1972, 1976), Feldstein (1972), and Diamond (1975). These competing considerations pose a tradeoff, especially because it often is supposed that necessities, consumed disproportionately by the poor, have relatively inelastic demands, and conversely for luxuries.

In addition to being widely taught, the Ramsey tax model and principles have provided the basis for extensive literatures on particular subjects, such as the taxation of capital, taxation and imperfect competition, and public sector pricing.[35] However, the foregoing analysis of optimal commodity taxation—which suggests that uniformity is optimal in the basic case without regard to demand elasticities or whether goods are disproportionately consumed by the rich or the poor—stands in sharp contrast to the leading principles of Ramsey taxation and thus calls into question results in the many literatures that build on the Ramsey model.[36] It is useful to set forth this tension briefly and explain why conflicting Ramsey principles are indeed inappropriate in the presence of an income tax.

Begin with the original Ramsey model in which individuals are assumed to be identical and the government's sole objective is to raise revenue with minimal distortion. When one allows for an income tax (linear or nonlinear)—one feature of which is the possibility of a uniform lump-sum tax or subsidy (which, unlike individualized lump-sum taxation, is feasible)—there is no need to rely on distortionary commodity taxation.[37] This result obviously does not depend on any special assumptions about the form of the utility function.

The reason that raising all revenue by uniform per capita taxes is problematic has to do with income distribution, for in a world in which individuals' abilities vary, the poor are hit hard by such a tax, whereas social welfare may be maximized when they receive net transfers. (Likewise, as noted, the simple Ramsey prescription arising from models that assume identical individuals favors commodity taxes that may fall most heavily on necessities.) When distributive concerns are incorporated, however, the analysis in subsection 4.2, drawing on the initial result of Atkinson and Stiglitz (1976), shows that differential commodity taxes also have no role in an overall optimal scheme (under simplifying assumptions examined in subsection 4.3). Although Ramsey rules modified for distributive considerations differ from the simpler prescriptions derived when individuals are assumed to be identical, they still generally involve adjustments that deviate,

(1971) and Stiglitz and Dasgupta (1971), and also Dasgupta and Stiglitz (1972) and Mirrlees (1972a) on how this result may differ when distribution is a concern.

[35] Regarding the latter, taxes or subsidies on private goods correspond to setting public sector prices above or below marginal cost, respectively.

[36] Not all Ramsey principles differ, notably, Corlett and Hague's (1953) argument (mentioned in subsection 4.3) that leisure complements (substitutes) should be taxed (subsidized) relative to other commodities.

[37] Following the discussion of normalizations in subsection 4.1, it is sometimes believed that a model with commodity taxation and no income taxation is equivalent to one that also allows linear income taxation. This, however, is incorrect because, as the discussion in the text (and the analysis in subsection 3.2) makes clear, an important feature of a linear income tax is that it permits a uniform lump-sum grant (or tax) g, which a system of pure, anonymous commodity taxation does not allow.

perhaps substantially, from uniformity—even when separability is assumed so that no differentiation is optimal with an income tax. For example, under Ramsey rules commodities consumed primarily by the rich (poor) should typically be taxed (subsidized) if inequality is sufficiently great and if distributive concerns are sufficiently important. But this result does not hold when an income tax is available. As explained above, any effect of commodity taxation regarding income distribution can better be produced directly, through the income tax, which undertakes redistribution in an across-the-board fashion. (Interestingly, the grant component of the income tax involves a uniform tax when only distortion is a concern, rendering commodity taxes unnecessary, whereas the grant is positive—a subsidy—in most simulations of an optimally redistributive income tax, under which commodity taxes are also unnecessary in the basic case.)

In sum, whether or not distribution is a concern, results derived in the original Ramsey framework, in which no income tax is available, fail to provide proper guidance in a world with an income tax. Accordingly, as Stiglitz (1987) suggests, Ramsey principles may be relevant in developing economies, in which income taxation may be infeasible (although he suggests that other modifications may be required), but not in developed economies. Likewise, as implied by Atkinson and Stiglitz's (1976) original paper and subsequently reinforced by Stiglitz (1987), various models and associated prescriptions based on the Ramsey framework are likewise poor guides when an income tax is available. See also Mirrlees (1994, p. 223) making a similar observation with regard to the analysis of public goods provision. Accordingly—and on account of space constraints—derivations of Ramsey tax principles are not covered here, and the reader is referred to the surveys cited at the outset of this subsection.

5. Other types of taxation

This section addresses additional major forms of taxation. As suggested in the introduction and subsection 2.2, most types of taxation can best be understood and their optimal use properly determined by examining them through interpretations or extensions of the model of optimal labor income and commodity taxation. In this fashion, one can obtain an integrated view of how different tax instruments should be used together to maximize social welfare, thereby achieving the revenue-raising and distributive objectives of taxation.

5.1. Capital taxation

In sections 3 and 4, income taxation referred to the taxation of labor income. In a static (one-period) model, the question of the optimal tax treatment of income from capital does not arise. Many forms of taxation, including the corporate tax as well as a standard income tax, do reach the returns to capital (savings), so extending the foregoing framework to capital income is important. Many dynamic issues are deferred to section 9; this section focuses on fundamentals.

5.1.1. Income versus consumption taxation

Following Atkinson and Stiglitz (1976, 1980), it is useful to begin by employing the model of commodity taxation to illuminate the difference between a classical, accrual income tax and a pure (cash-flow) consumption tax, taking advantage of the fact that consumption in different time periods can be conceptualized as consumption of different commodities. As subsection 4.1 explains, a uniform commodity tax at rate α (the same as a proportional consumption tax), which gives the budget constraint in expression (4.3), is equivalent to a linear tax on *labor* income at rate $t = \alpha/(1+\alpha)$, which gives the budget constraint in expression (4.4).

To introduce returns to capital and the possible taxation thereof, it is helpful to consider a two-period model wherein individuals work only in period 1 and consume in periods 1 and 2. (Period 1 can be thought of as an aggregate of one's working years and period 2 as retirement years.) Suppose further that there is only one type of commodity in each of the two periods, denoted c_1 and c_2. That is, we are considering a two-good version of the commodity tax problem in which the first commodity is period 1 consumption and the second commodity is period 2 consumption. Individuals' utility is $u(c_1, c_2, l)$. In this model, a pure labor income tax (equivalent to a uniform commodity or consumption tax) gives the budget constraint

$$wl(1-t) = c_1 + \frac{c_2}{1+r}, \tag{5.1}$$

where r is the interest rate (and $g = 0$ to simplify the exposition).

By contrast, a standard income tax is defined as a tax on both labor and capital income at the same rate. A common statement, referred to as the Haig-Simons definition, is that the income tax base equals consumption plus changes in wealth, the latter of which in the present model arises on account of earnings on first-period savings. More generally, it includes all returns to capital, such as interest, dividends, and capital gains (the latter determined in principle on an accrual basis)—and also allowing offsets for negative values (notably, interest payments and capital losses). When the income tax applies to labor and the returns to capital, the budget constraint becomes

$$wl(1-t) = c_1 + \frac{c_2}{1+r(1-t)}. \tag{5.2}$$

In comparing expressions (5.2) and (5.1), it is sometimes noted that a labor income tax is equivalent to an income tax that exempts the return to capital, and, given the aforementioned equivalence between a labor income tax and a consumption tax, that a consumption tax is likewise equivalent to an income tax that exempts the return to capital.[38]

[38] In similar spirit, the Haig-Simons definition is often rearranged to state that consumption equals income minus changes in wealth (net savings or dis-savings), an identity made use of in personal (cash-flow) consumption tax proposals that define the tax base as income minus all savings plus all dis-savings. See subsection 12.2.1.

It is also illuminating to rewrite expression (5.2) as

$$wl(1-t) = c_1 + \frac{1+r}{1+r(1-t)} \frac{c_2}{1+r}. \tag{5.3}$$

Because $(1+r)/(1+r(1-t)) > 1$ when $t > 0$, expression (5.3) indicates that a full income tax is equivalent to a labor income tax combined with a differential tax on second-period consumption. Dividing both sides of (5.3) by $1-t$, this budget constraint can also be written as

$$wl = \frac{1}{1-t}c_1 + \frac{1}{1-t} \frac{1+r}{1+r(1-t)} \frac{c_2}{1+r}. \tag{5.4}$$

Expression (5.4) indicates that a standard income tax is also equivalent to a differential commodity tax scheme under which second-period consumption is taxed at a higher rate than is first-period consumption.[39]

5.1.2. *Capital taxation more generally*

A standard income tax can be understood as a special case of a labor income tax combined with a supplemental tax on second-period consumption. To generalize, one can let t_r denote the tax rate applied to the return to capital, r, in which case expression (5.2) becomes

$$wl(1-t) = c_1 + \frac{c_2}{1+r(1-t_r)}. \tag{5.5}$$

When $t_r = 0$, we have a labor income tax or a pure (undifferentiated) consumption tax, and when $t_r = t$, we have a standard income tax. But we may also consider schemes under which t_r may take on any value, positive or negative. The choice between a standard income tax and a consumption tax thus poses a particular slice of the question of the optimal level of t_r. (Note as well that wealth taxes are equivalent to supplemental taxes on second-period consumption.[40])

The analysis in subsection 4.2 indicates that, when labor is weakly separable in the utility function, so we can write $u(v(c_1, c_2), l)$ (and other qualifications noted in subsection 4.3 are inapplicable), no differentiation is optimal, so t_r should equal zero. This means that a consumption tax is superior to an income tax and, for that matter, to any nonzero tax or subsidy on capital income. See Atkinson and Stiglitz (1976,

[39] The formulations in the text examine proportional income taxes. With nonlinear taxes, one could state equivalences with regard to marginal rates.

[40] An ex post wealth tax (i.e., a tax on savings plus interest, available for consumption in period 2) or an ex ante wealth tax (i.e., a tax on period 1 savings, which equal $wl(1-t)-c_1$) could be set at the rate $t_r r/(1+r)$. In either case, the result would be the same as that from supplementing a labor income tax at rate t with a capital income tax at rate t_r.

1980) and Stiglitz (1987).[41] The intuition is that one can achieve any degree of redistribution by adjusting the rate schedule, so the only remaining question concerns efficiency, as in the original commodity tax problem.[42] That is, we can ask whether an individual of a given earnings level should be taxed relatively more or less depending on whether more income is allocated to first-period or second-period consumption. In this basic case, neutrality is optimal because it avoids an additional distortion (of the intertemporal pattern of consumption), whereas differentiation would not help to offset the preexisting distortion (of labor supply).[43] This benchmark facilitates the analysis of reasons for departure from the zero-tax result, including arguments favoring capital income taxation (although it is unlikely, except on administrative grounds, that the optimal level of t_r would precisely equal t, as under a standard income tax).

One reason for departure is nonseparability. For example, if higher consumption (viewed here as an aggregate in each period) in period 1 enhanced the value of leisure whereas consumption in period 2 has no effect, it would be optimal to subsidize savings relative to first-period consumption. Another is myopia—see, for example, Laibson (1998)—which also may favor savings subsidies.[44] A recent body of work, surveyed by Golosov, Tsyvinski, and Werning (2007), indicates that capital taxation may be efficient to counter socially excessive precautionary savings. Additionally, in a general equilibrium setting in which wages are not given, capital taxes or subsidies may be optimal if they favorably influence the distribution of pre-tax income through the effects of changes in the capital stock on wage rates. See Stiglitz (1985b). (Compare the general equilibrium effects noted in subsection 3.5.2.) In this instance and more broadly, when the government cannot directly control the capital stock through debt or other policies, taxation or subsidization of capital serves as a substitute instrument.[45] Separate argu-

[41] The result that no capital taxation is optimal also arises asymptotically in models with infinitely-lived individuals, see Judd (1985), Chamley (1986), and the survey in Auerbach and Hines (2002), although these analyses are in a Ramsey setting, on which see subsection 4.4.

[42] Accordingly, the notion that consumption taxes are less redistributive than standard income taxes because the rich have more savings and thus more capital income is not emphasized here; adjusting the tax schedule to allow a distribution-neutral comparison clarifies the analysis of intrinsic differences between the two types of taxation. In practice, it is notable that the United States, which relies primarily on the income tax for redistribution, is generally viewed as engaging in less redistribution than many European countries, most of which rely heavily on a VAT, a form of consumption taxation.

[43] As Feldstein (1978) emphasizes, the extent of intertemporal distortion is not indicated by the change in savings but instead by the effect of differential taxation on consumption across periods. For example, even if savings were unaffected, it is still true that c_2 falls relative to c_1 as t_r increases.

[44] On taxation and saving more generally, including behavioral theories, see Bernheim (2002). On myopia, capital taxation, and labor supply, see Kaplow (2006a).

[45] For further exploration using overlapping generation models, see, for example, Atkinson and Sandmo (1980), Atkinson and Stiglitz (1980), Ordover and Phelps (1979), and Stiglitz (1985b). If individuals differ not only in earning ability but also in their ability to invest successfully—and if, moreover, capital markets are imperfect so those most productive at investment do not manage others' savings—additional subtle adjustments may be optimal. See Stiglitz (1985b).

ments that favor consumption taxation over income taxation focus on administrative grounds, some related to issues explored in section 9.[46] See subsection 12.2.1.

5.1.3. Corporate taxation[47]

The corporate income tax (in its classical, unintegrated variant) is levied on equity investment undertaken in the corporate form. Specifically, corporations subject to it pay an income tax on their earnings, and individual taxpayers pay a further round of tax under the personal income tax on dividend distributions. (Individuals also pay tax on interest receipts, but interest payments are deductible to the corporation.) By contrast, investments through sole proprietorships, partnerships, and certain types of corporations are not subject to an entity-level tax; instead, income is attributed to owners who are taxed accordingly.[48]

The corporate tax adds a further layer to the foregoing analysis of capital taxation. Just as one can determine whether capital taxation is efficient by holding the distribution of income constant, so one can assess intrinsic features of the corporate tax most directly by considering changes in its level as part of a reform that keeps the level of capital taxation constant. Viewed in this light, the central feature of the corporate income tax is that capital invested in certain legal forms is subject to a higher level of tax than capital invested in other forms. Moreover, as noted, because interest is deductible, the corporate tax only applies to corporate equity. Such differential taxation tends to distort investment decisions, in the present context by discouraging operation in the corporate form, by encouraging the use of debt rather than equity, and perhaps also (see below) by discouraging dividend distributions, in each case relative to the levels that would be chosen for nontax reasons.[49]

A natural question to consider is: Why tax corporations per se?[50] Most analysts, who emphasize that the burden of the corporate tax is ultimately borne by individuals, are

[46] It is worth noting that the core difference between income and consumption taxation is less significant regarding existing income tax regimes than may appear to be the case. First, as suggested in portions of sections 9 and 12, actual income taxes exempt or tax at a lower rate much of capital income. (Notably, human capital is largely taxed as it would be under a pure consumption tax; likewise for retirement savings. Imputed income from owner-occupied housing is exempt, dividends and capital gains may benefit from preferential rates (including the exclusion of capital gains at death), and the realization requirement provides substantial deferral on much remaining capital income.) Second, as subsection 9.2.2 indicates, the tax on capital income falls primarily on the riskless return, well below the total return on equity.

[47] See generally Atkinson and Stiglitz (1980), Auerbach (2002), Bradford (1980, 1986), Graham (2003), Gravelle (1994), King (1977), and McLure (1979).

[48] Rules that vary across jurisdictions and over time determine which entities are subject to the corporate tax. In the United States, most large, widely-held entities (and many others) are covered.

[49] Another effect of corporate taxation is that it induces avoidance behavior—including in recent times the increasingly creative use of financial instruments to issue equity-like securities that will be treated for tax purposes as debt—and governmental regulatory responses that themselves consume resources.

[50] Justifications and problems that arise in an international setting are not considered here.

skeptical that good reasons exist. One justification is that the corporate tax prevents avoidance of the individual income tax on capital, for in the absence of a corporate tax, individuals could invest in corporate form and defer taxes on capital income until they withdraw their funds, capturing the interest on tax that would otherwise be due in the interim. However, most proposed reforms involve methods of integration under which such deferral would not be possible.[51] It is also suggested that corporations benefit from limited liability and thus should be taxed; however, the argument is a non sequitur (prices should equal marginal costs, which here may be near zero, not benefits), the corporate tax obligation is not directly related to any such benefits, and other limited liability entities are not subject to the tax. Various additional theories, based on different governmental benefits or other grounds, have been offered, but few relate closely to the form of the corporate income tax and most apply in principle to entities not subject to it.

Regarding distortion, the seminal contributions by Harberger (1962, 1966) present a general equilibrium model that, among other results, shows how the tax is likely to be borne by all capital, not just corporate equity. The basic point is that, in equilibrium, all forms of investment must offer the same after-tax rate of return; with differential taxation, this condition implies differences in before-tax returns, which are the source of distortion. Accordingly, the corporate tax imposes welfare costs if there are nontax reasons—perhaps relating to agency problems, asymmetric information, and costs of financial distress—that some firms would find it efficient to employ the corporate form, to use equity rather than debt, and to distribute rather than retain earnings.[52] For estimates, see, for example, Goolsbee (1998, 2004), Gordon and MacKie-Mason (1994), Gravelle (1989), Gravelle and Kotlikoff (1989), MacKie-Mason and Gordon (1997), and U.S. Department of Treasury (1992).

An issue that has proved perplexing concerns dividend distributions. The new or tax capitalization view (contrasted with the so-called traditional view) holds that distortion is limited to contributions to new equity because the effect of the corporate tax on preexisting equity is capitalized into share prices in the first instance. See Auerbach (1979), Bradford (1981), and King (1974) developing the new view, and subsequent analysis and surveys in Auerbach (2002), Gravelle (1994), Poterba and Summers (1985), and Zodrow (1991). Under this view, repealing the corporate tax would confer a windfall on

[51] Methods include treating the corporation as a pass-through entity, like other entities; giving corporations a deduction for dividends paid like their deduction for interest; and giving shareholders a credit for corporate taxes paid or an exclusion for dividend income. See, for example, American Law Institute (1989, 1993), McLure (1979), and U.S. Department of Treasury (1992). Most OECD countries provide some degree of integration, usually providing relief at the shareholder level. See, for example, Messere, de Kam, and Heady (2003).

[52] Were it not for nontax costs or legal limitations, firms might, for example, use exclusively debt to finance incremental investments, thus avoiding the marginal distortion caused by the corporate tax. See Stiglitz (1973).

previously invested capital—although transition provisions might avoid this effect.[53] It is disputed whether the new view is empirically valid. Most analysis of the effects of the corporate income tax, particularly regarding firms' dividend policy, is confounded by the uncertainty about why corporations pay taxable dividends in the first instance, especially when share repurchases accomplish very similar results but do not subject shareholders to the tax on dividends (but only to taxes on capital gains that often would be lower).

5.2. *Transfer (estate and gift) taxation*

Many jurisdictions impose taxes on voluntary transfers, either nominally on the gifts and estates of donors or, through inheritance or accessions taxes, on the receipts of donees. Whether such taxation, usually limited to large transfers, is appropriate, should be expanded, or should be repealed has proved controversial. See, for example, Aaron and Munnell (1992) and Joint Economic Committee (1999). A closely related issue concerns the treatment of voluntary transfers in income tax systems: Generally, there is no deduction to the donor and no inclusion by the donee (although some, notably Simons (1938), advocate such inclusion). Ordinarily, all that matters will be the aggregate net tax or subsidy on transfers, so the analysis here will proceed accordingly, not distinguishing among these forms of transfer taxation.

Following Kaplow (2001), transfers (hereinafter, generically referred to as gifts) can be analyzed as a specific form of consumption. A donor's utility is given by $u(c, c_\gamma, l)$, where c refers to expenditures on own-consumption and c_γ to expenditures on gifts to others. The donor's budget constraint can be depicted as

$$wl - T(wl) = c + (1 + t_\gamma)c_\gamma, \tag{5.6}$$

where t_γ is a differential tax or subsidy on giving.[54] The question is the optimal sign and magnitude of t_γ. As with commodity taxation generally and capital taxation, the present analysis—unlike much policy debate—does not consider revenue-raising or redistribution to be central to understanding transfer taxation. The reason, as before, is that the income tax can be adjusted in a revenue- and distribution-neutral fashion, leaving only the efficiency effects that are intrinsic to the specific form of differential taxation. The question presented is whether, at a given level of income, a donor should be taxed relatively more or less on account of giving an additional dollar to a donee rather than spending it on own-consumption.

The answer might appear to be the same as that for the general differential commodity tax problem considered in section 4. With weak separability—if we can write the

[53] Viewing the benefit to old equity as a pure windfall is subject to doubt because anticipation of such relief from corporate tax would partially offset the existing distortion, especially given the long delay between contemplation of integration, which has been ongoing, and its ultimate enactment. See subsection 9.3.

[54] It is often imagined that much giving, including all bequests, is from savings; to the extent this is the case, gifts can be embedded in the multi-period model considered in subsection 5.1 on capital taxation.

donor's utility function as $u(v(c, c_\gamma), l)$—we know that $t_\gamma = 0$ is optimal. One might imagine that, instead, giving is a substitute for leisure (i.e., the less own-consumption, the less valuable is leisure because one has less funds to spend on leisure activities), justifying a subsidy. Or the relationship may be the opposite, for example, if giving more to grandchildren increases the value of time spent with them (although if it is in retirement, labor supply may be unaffected).

Such an analysis of giving, however, is incomplete because it ignores the effects of gifts on donees. For each person who is a potential donor, suppose that there is a single potential recipient, whose budget constraint is

$$wl - T(wl) + \gamma = c, \tag{5.7}$$

where the γ in (5.7) equals c_γ in (5.6), the amount given by the donor, and c in (5.7) is total consumption by the donee. Giving by donors thus involves two sorts of externalities. First, there is a positive effect, that on the donee's utility. Even an altruistic donor (see subsection 10.4) considers only the effect of the donee's gain on the donor's own utility, whereas an SWF will also count the utility gain to the donee per se. This suggests a basis for subsidization. Second, gift receipts produce an income effect on donees, leading to a reduction in labor supply.[55] With a preexisting income tax, this involves a negative externality to the public fisc. (A contrary effect arises to the extent that gifts relax liquidity constraints, such as by enabling investments in human capital or entrepreneurship.[56]) Depending on the relative magnitude of these effects, a subsidy or tax may be optimal.

Although individuals' motives are often irrelevant in economic analysis (for example, it usually will not matter why individuals prefer a particular mix of vacations and home-based leisure activity), motives are important in analyzing voluntary transfers. Whether a gift is motivated by altruism, various forms of warm-glow giving (see Andreoni (1990)), exchange (see Cox (1987) and Bernheim, Shleifer, and Summers (1985)), or accident (notably, accidental bequests due to imperfect annuity markets) may have important effects on how taxes or subsidies affect giving behavior and also on how any particular giving pattern affects donors' and donees' utility. See Kaplow (2001).

A natural extension to consider involves charitable giving.[57] Although often viewed as a subject in its own right, it clearly is a species of voluntary transfer, and the foregoing model and analysis is largely apt. Donors would be treated in the same fashion, and charitable organizations can be seen as conduits for individual donees (directly, such as when funds are dispersed to the poor, or indirectly, such as when medical research is produced that ultimately benefits victims of disease). The case for subsidy may be

[55] See, for example, Holtz-Eakin, Joulfaian, and Rosen (1993), Imbens, Rubin, and Sacerdote (2001), and Joulfaian and Wilhelm (1994).

[56] See, for example, Blanchflower and Oswald (1998), Cox (1990), and Holtz-Eakin, Joulfaian, and Rosen (1994a, 1994b).

[57] For a survey, see Andreoni (2006).

greater with many charities on account of the production of public goods; that is, although all pure giving produces a positive externality, the externality may on average be larger with certain charities than with others or than in the case of direct gifts to particular individuals. This may help to explain why subsidies to charitable giving, such as through an income tax deduction, are currently employed.

5.3. Social security taxation[58]

In many countries, payroll (labor income) taxes are levied on individuals (or, equivalently, on their employers) to fund retirement insurance, referred to in the United States as social security.[59] At one extreme, if there was no linkage whatsoever between individuals' payroll tax payments and their own retirement benefits, the taxes could be analyzed precisely as before. At the other extreme, if an individual's tax payments funded the equivalent of an individual account, earning the market return—and, moreover, if individuals would have saved at least as much in any event—the system would have no effect at all.

In reality, tax-benefit linkages exist but are complex. In the United States, some individuals receive no marginal benefits for their tax payments (young workers, very-low-income workers, some second earners), some pay a negative net tax (because own plus spousal benefits exceed in present value the marginal tax cost), and many individuals pay positive taxes net of benefits at widely differing rates that vary over their working lives. The divergence arises because, on average, many currently working cohorts will receive benefits less than taxes (whereas those first covered by the system received benefits significantly in excess of taxes)—see, for example, Leimer (1994)—and because there is substantial intracohort redistribution, including direct rich to poor redistribution through benefit formulas and supplemental assistance for the poor, offsetting redistribution because retirement annuities are more valuable to those with greater longevity (who tend to be higher-income individuals), and significant transfers among different family units (notably, to married couples having a spouse with no or low earnings). See, for example, Blinder, Gordon, and Wise (1980), Coronado, Fullerton, and Glass (1999), Feldstein and Samwick (1992), and Liebman (2002). Any net tax or subsidy at

[58] See generally Diamond (2002, 2003, 2004), Feldstein (2005), and Feldstein and Liebman (2002a, 2002b). These and other overviews of social security address many important issues beyond the scope of this chapter (with its focus on taxation per se), notably including funding (various forms of pre-funding versus pay-as-you-go systems, particularly with regard to effects on national savings), investments (investment mix, such as in equities versus government bonds, and private versus public control), how benefit rules affect retirement decisions, annuitization of benefits at retirement, intergenerational redistribution and risk-sharing, political economy considerations relating to benefit levels, and how current social security surpluses affect overall government deficits.

[59] This brief subsection focuses on income replacement, but many systems also fund medical care, disability insurance, and unemployment insurance.

the margin operates, in principle, like a pure tax on labor income. Accordingly, redistribution through social security is not (as a first approximation) qualitatively different from redistribution through the tax and transfer system.[60]

Two major qualifications are important. First, given the complexity of the formulas relating current taxes to future benefits, that benefits are far into the future and have contingent values, and that there is considerable uncertainty concerning future benefit levels (given that most systems are not in long-term fiscal balance), there is room for substantial heterogeneity in beliefs and outright misperception about tax-benefit linkages. See, for example, Dominitz, Manski, and Heinz (2003). It is suspected that many, especially younger workers, may underestimate the marginal benefits that accrue as they work, so that the labor supply effects of payroll taxes are greater than they would be if benefits were fully appreciated.

Second, a major rationale for social security is that forced savings is beneficial on account of individuals' myopia.[61] Many retire with few other assets, even though the social security replacement rate is significantly below plausible targets for optimizing life-cycle behavior. It may be that many individuals, even if they understood the benefits associated with the taxes they pay, would give such benefits little weight.[62] This factor may also seem to indicate that the effect of payroll taxes on labor supply is more analogous to that of simple taxes on labor income than to that of voluntary personal retirement contributions. Myopic individuals' behavior may be different, however, when one takes into account the influence of their myopia on savings decisions as well as labor effort. See Kaplow (2006a).

5.4. State and local taxation[63]

The central difference in analyzing taxes imposed by subnational jurisdictions is due to taxpayer mobility. In the perfect-competition version of Tiebout's (1956) model, individuals sort themselves into homogenous jurisdictions, each of which provides the desired public goods funded by benefit taxes, which, given the presumed homogeneity, would be uniform lump-sum (head or poll) taxes. In such a world, the payment of taxes to finance local public goods would be analogous to consumers' payments of prices to purchase private goods. There would be no distortion and, relatedly, no redistribution.[64]

[60] One difference is that net transfers through social security depend on lifetime income, although the income tax and transfer system could, in principle, as well. See subsection 9.6.

[61] Other rationales that do not have the same implications include asymmetric information and other shortcomings in the annuity market (see, for example, Brown, Mitchell, and Poterba (2002)), the Samaritan's dilemma (see Buchanan (1975)), and a desire to redistribute based on lifetime income.

[62] If so, social security may not significantly displace private savings for such individuals.

[63] See generally Inman and Rubinfeld (1996), Mieszkowski and Zodrow (1989), Musgrave and Musgrave (1989), Oates (1972, 1999), Rubinfeld (1987), Scotchmer (2002), Wildasin (1986), and Wilson (1999). An important set of issues ignored in this subsection concerns horizontal relationships across taxing jurisdictions, the issues being analogous to those considered in subsection 5.5 on international taxation.

[64] An exception would arise to the extent that redistribution itself is a local public good. See Pauly (1973).

In fact, smaller jurisdictions rely on other forms of taxation. In the United States, for example, localities heavily use property taxes and states primarily employ sales and income taxes. Because there is significant, even if imperfect, mobility and because of local political forces, the distribution of benefits and of taxes still tends to be somewhat aligned. To the extent that the coincidence is incomplete, there may be redistribution and corresponding distortion. See section 7 (which implicitly refers to a national government's provision of public goods). However, redistribution may not occur because tax-benefit divergences may be capitalized into land prices, as suggested by Hamilton (1976), although under the new view of the property tax developed by Thomson (1965), Mieszkowski (1972), and Aaron (1975), the tax is borne by owners of capital and capitalization may not occur.

Additionally, even if benefits equal taxes, this equivalence will tend to hold on average rather than at the margin. Thus, a worker contemplating additional labor supply may not expect to benefit more from public goods, in which case labor income taxes (and, relatedly, sales taxes) will tend to have effects like those analyzed previously. Similarly, property taxes will tend to distort investment in housing and other structures.

5.5. *International taxation*[65]

Most international issues in taxation concern how the effects of capital taxation differ in an open economy. In the often-studied limiting case, capital is perfectly mobile and the taxing jurisdiction is small.[66] A fundamental distinction arises between taxation of capital supply (saving) by residents, which may be invested domestically or in foreign jurisdictions, and taxation of capital use (investment) by location, which may be from domestic or foreign investors. In a closed economy, there is ordinarily no difference between the two because the incidence of a tax is unaffected by the side of the market on which it is nominally imposed.

Taxes on investment in a home jurisdiction—so-called source-based taxation—lead to a reduction in investment until the point at which the after-tax return is as high as for investments elsewhere. Hence, the incidence of such a tax falls on domestic labor and other fixed factors, not on domestic capital. From the home country's perspective, such taxation is inefficient: It distorts production without producing any unique benefit, such as taxing domestic capital (if that is desirable) or extracting any benefits from foreign investors.[67] Accordingly, although consistent with "capital import neutrality,"

[65] See generally Dixit (1985), Gordon and Hines (2002), and Slemrod (1988). Of necessity, this subsection omits many substantial topics, including avoidance and evasion issues that are prominent in an international setting, the corporate tax and its application to multinational enterprise, uncertainty and investor portfolio diversification, the interaction between tax policy and limits on capital mobility (including government actions that tend to offset capital flows), transition issues (accentuated by the fact that a capital levy would in part be borne by foreigners), and concerns about other countries' retaliation and the scope for international agreements.

[66] Similar analysis is applicable as well to capital taxation by subnational jurisdictions.

[67] The inefficiency of such taxation is an implication of the production efficiency result described in note 34.

source taxation is not generally favored by economists writing on international taxation, although such taxation is typically employed by developed countries.

Contrariwise, a so-called residence-based tax, equally applicable to residents' investments at home and elsewhere, reduces their return to saving but has no effect on domestic investment.[68] The analysis of such a tax is similar to that in subsection 5.1 on the taxation of capital generally (the earlier model suppresses production by implicitly assuming constant returns and competitive pricing, so the focus is on the utility of investors). Residence-based taxation, consistent with "capital export neutrality," is also consistent with global production efficiency because the location of investment is not distorted. Developed countries in fact levy taxes on residents' capital income. In addition, they often provide a foreign tax credit or occasionally exemptions to avoid double taxation arising from source countries' taxation of the same investments. As Richman (1963) explores, however, it would seem to be in a taxing jurisdiction's interest to allow only a deduction for foreign source-based taxation, treating it as a cost of doing business. A credit or exemption leaves investors indifferent between investing at home or elsewhere, but the home treasury is not indifferent because shifting the marginal dollar elsewhere costs the domestic fisc. Note, however, that the foregoing traditional analysis of international taxation treats capital as uniform, whereas capital taxation regimes may also distort ownership and thus the efficiency with which intangible capital (intellectual property) is transferred within multinational firms. See Desai and Hines (2003).

If a country has market power, notably if it supplies or demands a large share of the global capital stock, its nationally optimal policies tend to differ, by analogy to optimal tariff analysis. A large net capital importer will wish to tax the inflow and a large net exporter benefits by taxing the outflow.

6. Taxation and transfer payments

As subsection 2.1 indicates, two of the main purposes of taxation are redistribution and raising revenue to finance public goods and services. Accordingly, some analysis of expenditures—on transfers and on public goods—is essential to a full understanding of taxation. Furthermore, as subsection 2.2 emphasizes, the optimal design of one part of the fiscal system depends on what other instruments are available and how they are to be used, so any analysis that focuses exclusively on a subset of the system is incomplete and potentially misleading. Accordingly, this section and the next explore the two main categories of government expenditures, transfer payments (this section) and public goods and services (section 7).[69]

[68] Allowing for both types of taxation, the result may well be specialization in simple models. All domestic saving may be invested locally or all local investment may be supplied by foreigners, depending on the relative levels of taxes on residents investing domestically, on residents investing in foreign jurisdictions, and on foreigners investing locally. See, for example, Slemrod (1988).

[69] For a more in-depth treatment of transfers, see Kaplow (2007b).

6.1. Optimal transfers

A substantial and growing literature addresses the design of transfer programs.[70] One might, however, regard such treatments as unnecessary, at least at an abstract level; after all, section 3 on optimal income taxation presents the optimal tax *and transfer* schedule, $T(wl)$. Viewing transfers as part of the optimal tax problem is the approach adopted in a book review by Diamond (1968) that predates the leading modern contributions on optimal income taxation, and it is advanced in Mirrlees's (1971) conclusion, but it has not been followed very directly in most subsequent work.

Drawing on the analysis in section 3, the most straightforward answer to the question of how best to design transfers, notably concerning the optimal level of transfer to individuals earning no income and the optimal phase-out rate, is that one should simply inspect the lower end of the $T(wl)$ schedule depicted in Figure 3.1. As noted, $-T(0)$ is the transfer g (which can be treated as the combined value of all transfer programs to those earning 0 income), and $T'(wl)$ is the net phase-out rate (combining the phase-out of all transfer programs with any positive or negative marginal income tax or other tax separately imposed) at any income level wl. In other words, each tax and each transfer program can be represented by its own schedule $T^i(wl)$, and we can let $T(wl) = \sum T^i(wl)$.[71] An immediate and obvious implication is that one cannot meaningfully ask what in principle is the optimal design of a particular transfer program (or of a particular aspect of the income tax schedule, such as the EITC in the United States), for all that matters is the aggregate schedule, not the shape of any particular component.

Simulations reported in section 3 suggest that optimal grants are fairly generous in a wide range of settings and that optimal marginal tax rates (phase-outs) are significant as well. For example, Stern's (1976) simulations for a linear income tax have a central-case grant equal to 34% of average income and an optimal tax rate of 54%, Slemrod et al.'s (1994) optimal two-bracket simulations for similar parameters have a similar grant and somewhat higher marginal rates at the low end (approximately 60%), and simulations for the nonlinear case generally feature marginal rates at the bottom that were at or near the highest and often fairly high in absolute terms. Recalling the intuition underlying these results for marginal tax rates (see the discussion of expression (3.9) in subsection 3.4), rates at the bottom collect revenue from most of the population but are inframarginal (and thus not distorting) with regard to them, they do not apply to an

[70] For surveys, see Atkinson (1987a) and Moffitt (2002).

[71] A complication is that some transfer programs have a so-called cliff or notch effect, such that when income reaches a certain point, a particular benefit is lost altogether; that is, some $T^i(wl)$ may be discontinuous, so $T'(wl)$ may not be defined at particular points. Another complication is that some transfer programs are subject to asset tests, which is to say, for example, that an individual must first consume all assets before becoming eligible for transfers. This formulation can act as a 100% tax on the principal of one's savings in low-income states of the world which, as Diamond (1968) explains, can be a substantial deterrent to savings. Some evidence suggests that asset tests indeed discourage savings. See Gruber and Yelowitz (1999) and Powers (1998); but see Hurst and Ziliak (2006).

extremely high density of the population (which tends to be at a maximum closer to the middle of the income distribution), and they involve little productivity and thus revenue loss per unit of work effort that is sacrificed.

The existing system of transfer programs in the United States has high aggregate marginal rates (consisting mostly of welfare phase-outs) near the bottom of the income distribution, indeed, even higher than seems likely to be optimal. Before the welfare reforms in the mid-1990s, Giannarelli and Steuerle's (1995) microsimulations found aggregate marginal rates averaging 75% or more near the bottom and that many faced rates of 100% or more. Post-reform, Sammartino, Toder, and Maag (2002) find average marginal rates of roughly 60%–70% near the poverty line and over 100% in a modest range just above the poverty line. Other studies report similar results.[72]

6.2. *Categorical assistance*

Transfer programs often are targeted at or are more generous to particular groups, usually individuals deemed less able to work on account of age, family configuration (single parent with young children), or disability. In this respect, transfer programs serve as ability-based taxation, as discussed in subsection 3.5.1. It is useful to extend that analysis by considering some special cases. First, suppose that it is possible to observe perfectly which individuals' abilities are below some low level, w°. Then that group can be given a high transfer g that would not be very costly to finance; g could be fairly low for everyone else without fear that such individuals would be destitute because, by assumption, they all can earn at least a minimal income. Relatedly, it would be optimal not to tax low levels of earnings in the group for whom $w \geq w^\circ$, thereby avoiding any labor supply distortion at the bottom of the group.[73]

More realistically, signals about ability will be noisy. Even though some features, such as age or certain disabilities, can be observed nearly perfectly at low cost, there will usually be differences in ability associated with these characteristics. And other traits,

[72] See, for example, Dickert, Houser, and Scholz (1994) and Wilson and Cline (1994) before welfare reform, and the post-reform studies by Acs et al. (1998), Gokhale, Kotlikoff, and Sluchynsky (2002), and Hepner and Reed (2004). These studies take into account the existence of the EITC: In the credit phase-in range, the EITC mitigates aggregate marginal rates for very low-income individuals, but in its phase-out range, aggregate marginal rates are pushed higher. Although the EITC has received substantial attention and is viewed by nonspecialists as an alternative to traditional welfare, it should be kept in mind that in fact it is roughly equivalent (for the relevant population) to stretching out the phase-out of welfare. (The marginal subsidy in the phase-in range has the same effect as reducing the phase-out rate of other programs to the same extent, and the EITC phase-out is equivalent to extending welfare phase-outs by raising their level and applying them to higher-income individuals, on account of the need to complete the phase-out that is prolonged by the EITC subsidy.)

[73] The usual argument for high marginal rates near the bottom of the income distribution is inapplicable when all earn above some minimum level because then it is more efficient to reduce the grant g than to apply a positive marginal tax at the bottom. This is the result in the optimal income tax literature, discussed in note 20, that a zero rate at the bottom is optimal, which as mentioned is inapplicable in the ordinary (noncategorical) case in which the lowest-ability individuals do not work.

including many other disabilities, cannot be observed perfectly. Accordingly, suppose that a low-cost signal makes it possible to divide the population into two groups, group L consisting mostly of individuals with very low ability and group H containing few such individuals. That is, by reference to the population density function $f(w)$ from the optimal income taxation model in section 3, the density $f^L(w)$ is heavily concentrated at low levels of w and the density $f^H(w)$ is very thin at the bottom.[74]

The analysis of the optimal nonlinear income tax provides the basis for making conjectures about the optimal tax and transfer schedule for each group. To begin, it seems plausible that $g^L > g^* > g^H$. To consider the shape of the optimal tax schedule for each group, consider the first-order condition (3.9), reinterpreted for the present case involving two groups in the manner suggested in subsection 3.5.1. For the more able group, focus on low levels of income. The first term will be notably higher than in the single-group version of the problem. The $1 - F$ component will be somewhat greater because almost everyone in the group will have higher incomes; more significantly, the f component in the denominator will be smaller, indeed, very small if the categorization is even moderately accurate. This suggests that the optimal marginal tax rate at low levels of income should be substantially higher than in the standard problem.[75] For the less able group, the opposite result seems plausible. The $1 - F$ component will, after extremely low levels of income, be substantially smaller than in the combined problem, and f will be much larger, favoring low marginal rates in this income range.[76]

Interestingly, existing welfare phase-outs tend to have the opposite character: When benefits are high, as they are for low-ability groups, aggregate phase-out rates are correspondingly high because there are more benefits being phased out. But it was just suggested that optimal aggregate (phase-out inclusive) marginal tax rates for such individuals may be low, even if that means that the substantial grant is not fully phased out until income reaches higher levels. For high-ability groups, benefits are low so there is little to phase out and, accordingly, phase-out rates are low. Yet the foregoing analysis explains that high marginal tax rates may nevertheless be optimal. This apparent deviation from optimality seems to be a product of focusing on transfer programs in a vacuum, as if they are subject to their own special requirements. Specifically, it tends to be assumed that, when transfers are granted, they must be phased out, and that the

[74] The analysis assumes that the traits determining the categorization are exogenous. However, providing more generous benefits to certain family units or to the disabled may affect incentives to marry, procreate, or avoid injuries, which in turn will influence the optimal degree of differentiation of treatment between categories.

[75] Some offset will be provided through the second term because $g^H < g^*$ and the existence of higher marginal tax rates at low income levels each implies that individuals at higher income levels will have somewhat higher marginal utilities of income and (for strictly concave SWFs) higher welfare weights. On this account, it may not be optimal to have g^H much below g^*. After all, with a steep phase-out, few individuals will benefit much from the grant in the higher-ability group; those who do will be individuals who are misclassified.

[76] As the previous note indicates with regard to the higher-ability group, some offset will be provided by the second term: Because of the more generous grant and low initial marginal rates, the welfare cost of higher payments by those with greater income will be less than otherwise.

phase-out must be complete at reasonably modest levels of income, lest welfare become too expensive and available to non-needy individuals.[77] And when there is little welfare to be phased out, there correspondingly is thought to be no need for high marginal tax rates. A virtue of incorporating the analysis of transfer programs into the optimal income tax framework is that potential errors that result from unintegrated thinking can be avoided.

6.3. Work inducements

There has been ongoing concern with getting welfare recipients to work. Optimal income taxation analysis, however, does not attribute significance to work per se.[78] Inefficient work disincentives are a byproduct of positive marginal tax rates, and this cost is factored into the analysis. Additionally, it was noted that a feature of the optimum in a unified system is that the lowest-ability individuals do not work. In a perfect categorical system, this would continue to be true of those with the least ability, for it is optimal to give them a generous grant despite its work disincentive effect. If those above a minimum level of ability can be identified perfectly, they will be induced to work because their grant, g^H, will be set very low. When categorization is imprecise, these results will be approximated imperfectly.

Although the analysis thus appears to be complete, it is interesting to examine schemes that might induce additional work effort. One might reduce transfers by the extent to which earnings fall below some target level, perhaps the income produced in a full-time minimum-wage job. The marginal return to work in the relevant range for one who could earn only the minimum wage would be double the wage (the earnings per se, plus a one-for-one reduction in the shortfall penalty) minus taxes and benefit phase-outs. Supposing that the latter aggregated to under 100%, the effective marginal tax rate would be negative, a net subsidy to earnings. If everyone subject to such a regime has the requisite ability such that, in an optimal scheme, they all would work at least at the target level, then it is unproblematic. However, if there are classification errors—notably, if some subject to the work requirement have a lower ability—then the foregoing analysis suggests that this scheme is not optimal. The possibly extremely low implicit grant level may well be too low, and marginal rates should be greater, possibly quite high rather than negative near the target. In any event, it is unlikely to be optimal for the marginal tax rate to jump on the order of 100 percentage points at the target income level.

Some work incentive schemes assume that hours as well as earnings are observable. As noted in subsection 3.5.1, earning ability might then be inferred (earning ability, the

[77] This assumption about transfers and phase-outs often characterizes not only political debate but also formal analyses, as reflected for example in Moffitt's (2002) survey.

[78] If there were positive externalities to work by the poor (setting a good example for one's children that has the effect of reducing future dependency and crime) or negative externalities (such as may flow from reduced supervision of children), the analysis would differ.

wage, is simply earnings divided by hours). When ability can be observed, the first-best can be achieved (individualized lump-sum taxes with zero marginal tax rates for everyone), and there is no role for work requirements. When hours are observed, however, ability is only conditionally observable; individuals who do not work at all do not reveal their type. The second-best optimum for this case similarly involves individualized lump-sum taxes (for those who do work) and zero marginal tax rates, although the extraction from higher-ability types is incomplete. See Dasgupta and Hammond (1980).[79] Existing programs premised on the observability of hours differ qualitatively.[80] Furthermore, as mentioned in subsection 3.5.1, hours are manipulable, so it is not obvious that the use of hours is feasible. See Moffitt (2002). With imperfect observability, the analysis of categorization in subsection 6.2 becomes applicable.[81]

6.4. *Cash versus in-kind transfers*

It is ordinarily supposed that cash transfers are superior to transfers in kind because individuals have different preferences and they tend to have better information than does the government about their own situations. Nevertheless, many transfers are given in kind, a practice that may sometimes be optimal for a number of reasons.[82] The poor may be myopic or otherwise unable to make wise spending decisions, a possibility strengthened by the fact that such infirmities may have contributed to their low earning ability. There may be externalities due to certain forms of consumption, such as if housing reduces crime or immunization prevents the spread of disease. There may be psychic externalities when taxpayers feel better knowing that the assistance they provide must be spent on food, shelter, medical care, or education. In addition, taxpayers may be subject to the Samaritan's dilemma, concerned about strategic imprudence by potential beneficiaries, a possibility that makes compulsory health and social insurance particularly appealing.[83] In-kind assistance can also direct aid to children in particular,

[79] See also related models by, for example, Diamond (1980) and Saez (2002) that focus on the participation decision.

[80] See, for example, Michalopoulos, Robins and Card (2005) on an experimental program in Canada that provides a large bonus to those working at least thirty hours. A more moderate work inducement, corresponding to that previously described in the text for the case in which only earnings are observed, takes the following form when hours are also observable. Let w° and l° denote the target wage and required labor supply and t the (flat) preexisting aggregate (inclusive of phase-outs) marginal tax rate below the target income level. Then, for $l < l^\circ$, disposable income available for consumption, c, is

$$\begin{aligned} c &= g + wl(1-t) - w^\circ(l^\circ - l), \text{ or} \\ &= [g - w^\circ l^\circ] + wl(1-\tau), \end{aligned}$$

where $\tau = t - w^\circ/w$. Note that wage subsidies are similar.

[81] There is also literature on the possible optimality of conditioning welfare on public employment. See, for example, Besley and Coate (1995) and Brett (1998).

[82] Many of the arguments are noted, for example, by Nichols and Zeckhauser (1982).

[83] See, for example, Bruce and Waldman (1991) and Coate (1995).

such as by providing medical care, education, or meals served at school. Furthermore, it may be possible to use in-kind assistance to help mitigate labor supply distortion in subtle ways by improving selection.[84]

7. Taxation and public goods

As noted previously, the revenue-raising objective of taxation makes analysis of the relationship between taxation and public goods central to a comprehensive understanding of taxation. Furthermore, because a substantial fraction of GDP consists of governmentally provided goods and services and because such provision itself has significant distributive implications, analysis of taxation and public goods is also important in understanding the redistributive function of taxation.[85] As one might expect, there are important interactions between the two subjects: For example, how much redistribution is optimally undertaken through the income tax will depend on the extent to which the public goods funded by income tax revenue benefit the poor. Thus, as in the case of other forms of taxation and of expenditures on transfers, it is necessary to examine income taxation and public goods together to know how each should be determined so as to maximize social welfare.

7.1. Distributive incidence and optimal redistribution

The distributive incidence of public goods is relevant to the optimal income tax problem because the optimal redistributive tax depends on individuals' utilities, which in turn depend on public goods. Specifically, when the SWF is strictly concave in individuals' utilities, the marginal social benefit of redistribution depends on the extent of differences in individuals' utility levels. Additionally, public goods may affect individuals' marginal utilities of consumption, which likewise affects this marginal social benefit.

Consider, for example, the case in which public goods are a perfect substitute for disposable income. That is, utility can be written as $u(c + b(G), l)$, where G denotes expenditures on public goods and b indicates the dollar-equivalent benefit. In this situation, public goods provide equal benefits measured in dollars (rather than utility) to everyone. Accordingly, the benefit of the public good is equivalent to a higher g, the uniform grant. Ignoring this effect of public goods would lead one to overstate (perhaps greatly) the marginal social benefit of redistribution.

[84] It is sometimes suggested that giving low-quality in-kind goods has this benefit. However, high-ability individuals generally mimic low-ability individuals by earning less, and given their lower earnings they may have similar preferences among goods to those of lower-ability individuals who earn the same amount. Hence, using in-kind provision of low-quality goods tends to be an inefficient means of improving screening. Compare Munro (1989). In contrast, in-kind provision, notably, of medical care, may helpfully select individuals of high need. See, for example, Blackorby and Donaldson (1988).

[85] The analysis in this section is applicable to any government expenditures on goods and services, regardless of whether they are public goods in the technical sense.

Now suppose that utility is additively separable in consumption, public goods, and labor: $u = v(c) + b(G) - z(l)$. Public goods provide equal benefits measured in utility (rather than dollars) to everyone. Measured in dollars, the value of public goods is given by the inverse of individuals' marginal utility of consumption, which depends on the curvature of v. For example, if $v(c) = \ln c$, then marginal utility is $1/c$, so the public good has a dollar value proportional to consumption. In this case, because of separability, the benefit of the public good has no effect on individuals' marginal utility of consumption and thus no effect on the optimal extent of redistribution through this channel, although raising everyone's welfare level by a constant amount (measured in utility) may affect the marginal social benefit of redistribution when the SWF is strictly concave.

In sum, knowing the distributive incidence of public goods as well as how public goods enter into individuals' utility functions (the two questions are closely related) is necessary to determine the optimal extent of redistributive taxation. Unfortunately, ascertaining distributive incidence empirically, especially for public goods like police protection and national defense, is notoriously difficult.[86]

Another question concerns how increasing government expenditures on some public good or service at the margin affects the desirability of income redistribution and, in particular, how this effect depends on the distributive incidence of the particular good or service. Suppose, following Kaplow (2006d), that the income tax was set optimally and that it became efficient (perhaps due to technological change) to supply more of some public good. Once that was done, what if any adjustment to the extent of redistribution would be appropriate? Clearly, the answer to this question will depend on how the public good is financed in the first instance. For example, if it were financed by taxing the poor (rich), more (less) redistribution is likely to be in order, but that would tell us little about how changing the public good per se affected the desirability of redistribution. Accordingly, it is useful to contemplate finance of the public good by a distribution-neutral (offsetting) income tax adjustment (see section 4), so the change in public good combined with its finance preserves the preexisting distribution. Under these circumstances, if additional redistribution then becomes desirable, it would be meaningful to say that changing the provision of a public good affects the desirability of redistribution.

One might conjecture that with distribution-neutral finance there is no further need to adjust the extent of redistribution. Indeed, when a public good is a perfect substitute for cash, such as in the first example above, and therefore is worth the same (dollar) amount to everyone, this conjecture can be shown to be valid. In the more general case, however, redistribution may become more or less desirable, through two channels. First, the reform package may affect the relative marginal utilities of the rich and the poor. (With a strictly concave social welfare function, changing relative utility levels would also matter, but the distribution-neutral tax adjustment keeps these constant.) Second,

[86] See, for example, Aaron and McGuire (1970), Musgrave, Case, and Leonard (1974), Reynolds and Smolensky (1977), and Ruggles and O'Higgins (1981) for the United States.

the reform package may affect the revenue impact of adjustments to redistributive taxation. After the hypothesized reform, raising marginal tax rates, for instance, may have different labor supply effects and, for a given labor supply effect, have different effects on revenue than before. Analysis of both channels is somewhat subtle, and no simple characterization has been obtained.

7.2. Distribution and distortion

The preceding subsection considers the implications of public goods provision for redistributive taxation. This subsection considers the reverse: how the second-best nature of redistributive taxation bears on optimal public goods provision. The first-best rule—the Samuelson (1954) rule—is that public goods should be provided until the point at which the sum of individuals' marginal benefits equals the marginal cost of provision. Two second-best caveats are standard.

First, Weisbrod (1968), Feldstein (1974), Drèze and Stern (1987), and others suggest that distributive weights be incorporated in cost-benefit analysis.[87] Second, a large literature following Pigou (1928) argues that the distortionary cost of finance should be accounted for in determining the optimal level of public goods.[88]

It turns out, however, that both qualifications arise from entangling choices of the extent of redistribution and of the level of public goods, decisions which can be separated in principle and in practice. Once the problems are unscrambled, the basic Samuelson test provides an appropriate benchmark for public goods provision. The reasoning closely parallels that used to analyze the inefficiency of differential commodity taxation in section 4. Indeed, some analysts have noted the analogy between the public goods and commodity tax problems.[89] Specifically, providing more (less) of a public good than is otherwise efficient is analogous to subsidizing (taxing) a particular commodity relative to other commodities. The analogy is even closer if one imagines a hypothetical public goods economy that employs Lindahl (1919) pricing, where a subsidy (tax) on the public good would entail lowering (raising) the "prices" consumers face for the public good just as commodity subsidies (taxes) on private goods entail lowering (raising) the prices consumers pay for them. Because the analysis is so similar to that of commodity taxation, it will only be sketched briefly.

Suppose that individuals' utility as a function of consumption c, public goods G, and labor effort l can be written as $u(v(c, G), l)$, which uses the assumption of weak separability of labor, just as in the analysis of commodity taxation.[90] (When this and

[87] One way of viewing this suggestion is that it responds to the objection to the Kaldor-Hicks hypothetical compensation test, see, for example, Little (1957), that standard cost-benefit analysis ignores distributive concerns.

[88] Pigou's argument was subsequently explored by Diamond and Mirrlees (1971), Stiglitz and Dasgupta (1971), and Atkinson and Stern (1974) in models of optimal taxation (in the Ramsey tradition; see subsection 4.4); and further developed in additional work, much of which is surveyed in Mayshar (1990), Fullerton (1991), and Ballard and Fullerton (1992).

[89] See, for example, Mirrlees (1976), Konishi (1995), and Kaplow (1996c, 2004).

[90] Both c and G could be interpreted as vectors, but the scalar representation is used to simplify exposition.

other assumptions are relaxed, qualifications parallel to those in subsection 4.3 are applicable.[91]) Again, the approach will be to identify the distribution-neutral (offsetting) adjustment to the income tax and transfer system, show that it does not affect labor supply, and determine when it generates a budget surplus or deficit.

The offsetting income tax adjustment for a marginal change in G is given by the marginal rate of substitution, v_G/v_c. To verify this, we need to consider whether this shift in the schedule T will be such that $\partial U/\partial G = 0$ for all types w and at every level of l that each type might supply. (This is a partial derivative because labor supply is being held constant; in the next step, it is shown that individuals indeed do not change labor effort when this tax adjustment is employed.) Thus, we consider

$$\frac{\partial U}{\partial G} = \frac{\partial U}{\partial v}(v_c c_G + v_G), \tag{7.1}$$

where c_G denotes the (here partial) derivative of c with respect to G. From the budget constraint (3.1),

$$c_G = -\frac{\partial T(wl, G)}{\partial G}, \tag{7.2}$$

where the notation $T(wl, G)$ is used to indicate how the tax schedule will be adjusted as G changes. If the tax adjustment is set equal to v_G/v_c, as suggested, then $c_G = -v_G/v_c$. Using this result and substituting (7.2) into (7.1) yields the conclusion that $\partial U/\partial G = 0$ for any given w and l.

Consider next whether individuals in fact would change their labor supply in response to a change in G financed by the specified adjustment to the tax schedule T. Just as in the analysis of commodity taxation in subsection 4.2, it should be apparent that, indeed, individuals of all types (w) would not change their labor supply. The reason is that expression (7.1) equals zero for any given l and hence for all l. Therefore (for each type w), if l^* was superior to all $l \neq l^*$ before G was changed, this will continue to be so afterwards because the utility at each and every l is unaltered by the change in G, when combined with the offsetting adjustment to T.[92]

Hence, government provision of a good or service when financed by a distribution-neutral (offsetting) income tax adjustment keeps everyone's utility (and hence the distribution of utility) constant and everyone's labor supply unchanged. It remains to determine how the government's budget is affected. But this is straightforward because the tax adjustment equals (at the margin) individuals' marginal rates of substitution. Total revenue due to the tax adjustment, therefore, is given by the integral of individuals'

[91] For example, improvements to public beaches and libraries, plausible leisure complements, should be undertaken to a lesser extent than that indicated by the Samuelson rule, whereas enhancing mass transit, a plausible labor complement, should be done to a greater extent.

[92] For a more formal derivation along these lines as well as one that differentiates the first-order condition for l with respect to G and uses the result to demonstrate that $dl/dG = 0$, see Kaplow (1996c, 2006d) and also Auerbach and Hines (2002).

marginal rates of substitution, so there will be a surplus (deficit) if the Samuelson rule is satisfied (fails). When the project passes the cost-benefit test, it is possible to rebate the surplus to make everyone better off (and when the project fails the test, a movement in the opposite direction will make possible a Pareto improvement).

It is immediately apparent why the two standard caveats are inapposite: One needs no special adjustment for distributive or labor supply effects because, on account of the use of a distribution-neutral income tax adjustment, there are none. Relatedly, if a public good were not to be financed by a distribution-neutral tax adjustment, the present approach is still warranted by the analysis of subsection 2.2, in particular the discussion of the virtues of using a two-step decomposition to separate the intrinsic effects of public good provision (hypothetically financed in a distribution-neutral manner to remove the pure effects of redistribution) from those of redistribution per se.

This view of public provision, which contrasts sharply with that in the literatures noted previously, is associated with an emerging body of work.[93] Hylland and Zeckhauser (1979) were the first to use offsetting tax adjustments to show that distributive incidence should be ignored. Christiansen (1981) and Boadway and Keen (1993) show that the simple cost-benefit test for public goods provision is correct if one assumes that the income tax is set optimally; they take advantage of the fact that, when at the optimum, the marginal benefit of additional redistribution equals the marginal cost of additional labor supply distortion, so marginal adjustments to the tax system have no net effect on social welfare. Kaplow (1996c, 2004, 2006d) builds on Hylland and Zeckhauser's approach to advance the view that both distribution and labor supply distortion can be ignored with regard to a wide domain of government policy, notably including public goods.[94] As should be apparent from the present analysis (and the analogous argument in section 4 on commodity taxation), when a distribution-neutral (offsetting) tax

[93] The reconciliation between previous work finding that distortion is associated with public goods provision and the present result is that the other work tends to find distortion when the combination of public good and tax adjustment in the policy experiment under consideration result in an increase in redistribution—yet the distributive benefit is disregarded (in a manner that may nominally be justified by the use of representative-agent models, which ignores that the motivation for employing distortionary taxes like the income tax rather than a uniform lump-sum tax is precisely that individuals are heterogenous, so distribution matters). See the discussion of Ramsey taxation in subsection 4.4 and also Kaplow (2004). See also Allgood and Snow (1998), who show that much of the difference in leading empirical estimates of the marginal cost of funds and of redistribution can be attributed to subtle ways in which different authors' simulations implicitly change the level of effective lump-sum transfers and thus the extent of redistribution assumed to take place.

[94] See also Ng (2000) and Slemrod and Yitzhaki (2001). Slemrod and Yitzhaki's formulation for the optimal provision of public goods allows for adjustments to the cost (taxation) and benefit (public good) sides of a standard cost-benefit equation to take into account (on each side) effects on labor supply and on distribution. As the analysis in the text indicates, in the basic case with distribution-neutral finance, all of these adjustments cancel. Moreover, with non-distribution-neutral finance, the two-step decomposition suggests that all effects that do not cancel will be associated with a pure change in redistribution. In this setting, it aids in analysis and interpretation (see subsection 2.2 and Kaplow (2004)) to think of the distributive effects from the public good and from the method of finance as being netted (and the resulting labor supply effects will similarly offset in part), so one is left simply with a single cost and benefit associated with the change in the extent of redistribution.

adjustment is employed, features of the initial income tax and transfer system—notably, whether it is set optimally—are irrelevant.

7.3. Benefit taxation

The foregoing discussion of the use of distribution-neutral (offsetting) income tax adjustments to finance public goods may be used to illuminate the long-discussed concept of benefit taxation. See, for example, Musgrave (1959). The first observation is that this particular mode of tax adjustment is indeed a sort of benefit taxation, as the magnitude of a marginal adjustment equals marginal benefits. (Recall that, at each income level, the marginal tax adjustment equals individuals' marginal rates of substitution.) For discrete changes, the offsetting tax adjustment equals individuals' total benefits for the project. Surplus is included, as the total tax adjustment at any level of income equals the area under the implicit demand curve for the public good.

This tax adjustment, however, differs from prior understandings of benefit taxation in important ways. First, as just explained, the posited tax adjustment is equivalent to Lindahl (1919) pricing at the margin, but it is not equivalent for a discrete change. Thus, if marginal benefits are declining, the average rate of the offsetting tax adjustment would exceed the Lindahl price, which equals the marginal benefit at the final point of the increase in G. Second, the stated tax adjustment differs from many notions of benefit taxation because it is based entirely on the benefits of a public project without regard to its cost.

Various authors have proposed a number of candidates for benefit taxation, most of which differ from the present formulation in other ways as well. See, for example, Hines's (2000) proposal and his review of Lindahl pricing and related alternatives. Such work usually presents as its objective the derivation of a benefit measure that has certain properties in common with the market's pricing of private goods or that meets other a priori criteria.[95] However, the purpose of providing such a measure is not explained. By contrast, the distribution-neutral (offsetting) income tax adjustment is chosen here because of its usefulness in policy analysis. This benefit measure also has descriptive functions. For example, whether the median voter will favor a project depends on whether that voter's actual new tax obligation exceeds this hypothetical offsetting tax adjustment.[96]

Yet another reason for formulating a principle of benefit taxation is to determine the proper manner of financing government expenditures on goods and services. Yet when

[95] See, for example, Hines (2000) and the debate between Aaron and McGuire (1970, 1976) and Brennan (1976a, 1976b).

[96] Another feature of the distribution-neutral approach is that it renders moot concerns about the progressivity of benefit taxation, a subject that has received much attention. See, for example, Hines (2000) and Snow and Warren (1983). Whatever is the degree of progressivity (or regressivity) of a tax that is set equal to actual benefits, its distributive incidence is, by definition, precisely offset by that of the public good itself. Hence, changing the level of public goods, financed by such benefit taxation, has no distributive effect regardless of the shape of the tax adjustment viewed in isolation.

there is also a system of redistributive taxation in place, which itself may be freely adjusted, the purpose of isolating the benefit tax component is unclear. From some normative perspectives, such as a libertarian one, benefit taxation may be required and any redistribution may be deemed impermissible. Such an approach does require selection of a particular definition of benefit taxation, and it is also necessary to confront the difficult (some would say insurmountable) baseline question regarding the benchmark against which one measures the distributive incidence of the entire public sector (is the hypothetical alternative anarchy?).

8. Corrective taxation

Correcting externalities is a third major function of taxation. Subsections 8.1 and 8.2 review the analysis of corrective taxation when externalities are the only concern. Subsection 8.3 integrates the analysis of corrective taxation with that of income and commodity taxes to determine social-welfare-maximizing corrective taxation when revenue-raising and distributive concerns are also relevant.

8.1. Pigouvian taxes and subsidies

Among the contributions of Pigou (1920) was his diagnosis of externalities in terms of divergences between private and social costs or benefits and his suggestion of taxes and subsidies as a possible cure.[97] To analyze Pigouvian taxation, we can extend the model of commodity taxation from section 4, following Kaplow (2006c). As before, there are n commodities, $x_1, \ldots, x_n$. Corresponding variables $e_1, \ldots, e_n$ denote the total consumption of each commodity: $e_i = \int x_i(wl) f(w)\, dw$, for all i. Individuals choose levels of consumption and labor effort l to maximize their utility functions $u(x_1, \ldots, x_n, e_1, \ldots, e_n, l)$. Utility may have any relationship to the level of the e_i's; that is, the external effect due to each of the commodities may be positive, negative, or nonexistent.

The monetary equivalent of the marginal external harm associated with any commodity x_i is

$$h_i = -\int \frac{\partial U(w)/\partial e_i}{\mu(w)} f(w)\, dw, \tag{8.1}$$

where $\mu(w)$ is the Lagrange multiplier on individuals' budget constraints (4.1) for individuals of type w, signifying the marginal utility of income. (U is also expressed as

[97] Although it is conventional to focus on externalities alone, similar analysis is applicable to some informational and self-control problems. That is, if there is no externality but individuals underestimate the private benefit of some activity, a subsidy may be used to offset the divergence, or if private harm is underestimated, a tax may be helpful. See, for example, Gruber and Mullainathan (2005) on the possibility that cigarette taxes may improve the welfare of myopic smokers.

a function of w because the partial derivative may differ by type due to differences in consumption and labor effort.) The first term in the integrand, therefore, is the marginal effect of the externality on utility divided by the marginal utility of disposable income, which gives the marginal externality for a given type w, measured in dollars. Note that for positive externalities, $h_i < 0$.

Using this expression for marginal external harm, we can define first-best Pigouvian taxes and subsidies as a commodity tax vector $\{\tau_1, \ldots, \tau_n\}$ having the property that $(p_i + \tau_i)/(p_j + \tau_j) = (p_i + h_i)/(p_j + h_j)$, for all i, j. Notice that the definition does not require that $\tau_i = h_i$, for all i. The reason has to do with normalization, discussed in subsection 4.1: If all commodity taxes are raised or lowered in such a manner as to leave all price ratios unchanged, individuals' behavior will be unaffected—if the level of the income tax is also adjusted to produce the same effective disposable income. However, it is useful to think of a case in which the only commodity taxes are Pigouvian, in which event it is true that $\tau_i = h_i$, for all i.

8.2. Choice of instruments

Pigouvian taxes and subsidies not only constitute an important instrument for the control of externalities, but our understanding of them also helps to illuminate other government interventions that address externalities. For example, free immunizations against communicable diseases are akin to a Pigouvian subsidy that brings the price to zero, which would be optimal if the external benefit happened to equal the cost. However, if the external benefit is even larger and if the private benefit is less than other private costs (which may include inconvenience, modest pain, and the risk of adverse side-effects), one could employ a subsidy in excess of 100% of the direct cost or, as is commonly done, impose a regulation that requires immunization.[98] More broadly, in examining the choice among regulatory instruments, it is helpful to keep such connections in mind.

In many settings, particularly involving negative externalities, the Pigouvian tax prescription is not employed. Two alternatives with important similarities to Pigouvian taxation are legal liability and tradeable permits. A rule of strict liability requires the injurer to pay the victim for all harm imposed, which from the injurer's point of view is similar to a Pigouvian tax.[99] A Pigouvian tax has the advantage that, because victims do not receive compensation, they retain incentives to mitigate harm—although this is a disadvantage to the extent that the activities of the injurer and victim combined bear excessive costs, creating insufficient incentives to undertake the combined activities and excessive incentives to integrate if that would eliminate tax liability.

Tradeable permits have the familiar virtue that, like Pigouvian taxes, they result in cost-minimization because each purchaser equates the marginal cost of harm reduction

[98] If vaccinations are always completely effective, compulsory immunization of everyone is not only suboptimal, but dominated by the laissez-faire solution. See Brito, Sheshinski, and Intriligator (1991).

[99] When harm is uncertain, one can think of a Pigouvian tax imposed probabilistically—which is how the tort system typically operates—or that a tax equal to expected harm is imposed with certainty.

to the (common) market price of permits. Permits importantly differ from Pigouvian taxation because the overall quantity of the externality is not optimized—by polluters equating marginal cost to the tax rate, itself set equal to marginal harm—but rather is determined by fiat. Note, however, that in principle the government could adjust the number of permits until the point at which the market-clearing price equalled marginal harm at the given quantity.

A greater contrast is provided by command and control regulation, whether imposed directly (such as with technological requirements) or indirectly (such as through liability involving injunctions or negligence rules under which damages are only owed if standards are violated). The common objection is that the government decision-maker has limited information, for optimal regulation of this sort requires not only information about harm (which is also required to set a Pigouvian tax) but also about the costs of various technologies.[100]

An important qualification to the inefficiency of some forms of regulation, particularly when implemented through liability rules rather than government edict, is that injurers and victims may bargain to more efficient results. This important point of Coase (1960) is true even when no liability is imposed for externalities, for a victim could pay an injurer to abstain from harm-causing activity if the cost of the harm exceeded the injurer's benefit from the activity. This solution, of course, tends to be infeasible in large-numbers cases—such as with industrial pollution and externalities associated with automobiles. Furthermore, even when bargaining may be feasible, asymmetric information interferes with efficiency, and accordingly there may be benefits of legal rules that more closely mimic Pigouvian taxes.

The choice of instruments problem is a good deal more complex than the foregoing suggests, involving comparisons between the government's and private parties' (including victims') information, concerns about the ability of injurers to pay for harm (especially large harms that may occur with low probability), considerations of administrative costs, and other factors. See, for example, the *Handbook* surveys on environmental regulation by Bovenberg and Goulder (2002) and Revesz and Stavins (2007) and that on tort liability by Shavell (2007). Nevertheless, it is worth emphasizing that much analysis of various forms of regulation is illuminated by the comparison with the straightforward principles of Pigouvian taxation. Likewise, it is useful to take advantage of this relationship when considering how the problem of controlling externalities interacts with other second-best considerations related to the other functions of taxation, the topic of the next subsection.

[100] Weitzman (1974) argues that setting quantities (regulation) may be superior to setting prices (taxes) when there is uncertainty, but his result arises largely because he rules out feasible instruments, notably a nonlinear Pigouvian tax schedule under which the marginal tax rate equals the marginal expected harm, and because he assumes that the linear instruments could not be adjusted over time, even when observable behavior revealed errors. See Shavell (2007).

8.3. Distribution and distortion

The standard Pigouvian prescription that taxes should be set equal to the marginal external harm applies in a world that is first best, other than for the externalities being corrected. Just as section 7.2 considered whether and how one should deviate from the Samuelson rule for public goods when there is a concern for distribution and labor supply distortion in a world that is second-best with regard to redistributive taxation, so we can ask whether these two considerations call for modification of the Pigouvian prescription. Distributive concerns are often expressed, such as in the argument that a gasoline tax, which may be used to internalize pollution and congestion externalities from driving, may be regressive and thus less desirable on this account.[101] Labor supply distortion has received substantial attention in recent literature.[102] Initially, some thought environmental taxation might produce a "double dividend"—both correcting an externality and also raising revenue without distortion, permitting reductions in distortionary income taxation—and a substantial subsequent literature has suggested that the problem is more complicated and, as it turns out, environmental policies may exacerbate the preexisting labor supply distortion due to income taxation.

It would seem, however, that in light of the previous analysis addressing these same two concerns in the contexts of differential commodity taxation (section 4) and public goods (subsection 7.2), one would expect, under similar simplifying assumptions, that first-best Pigouvian principles provide an appropriate benchmark for analysis—and that the qualifications to this conclusion (section 4.3) would be largely the same as in the previous settings. This indeed is the case, as developed in Kaplow (1996c, 2004, 2006c).[103]

One way to view the intuition is to return to the model of differential commodity taxation, wherein (with weak labor separability) uniformity was optimal. The intuition was that differential commodity taxation distorted consumption choices without reducing the distortion caused by redistributive taxation; hence, differential taxation is inefficient. Specifically, eliminating (or reducing) differentiation, with a distributively offsetting adjustment to the income tax schedule, results in a Pareto improvement. Extending that logic to the present case with externalities, the relative prices that result in no distortion are no longer ones with no differential commodity taxation but instead are consumer prices that reflect the full extent of any externalities, as in the definition of first-best Pigouvian taxes and subsidies in subsection 8.1. Hence, making this

[101] See, for example, Casler and Rafiqui (1993) and West (2004).

[102] For a survey and a collection of literature, see respectively Bovenberg and Goulder (2002) and Goulder (2002).

[103] See also Pirttilä and Tuomala (1997) and Cremer, Gahvari, and Ladoux (1998). The reconciliation between this conclusion and that in much of the literature that does find additional distortion associated with greater environmental regulation or particular environmental policies is, just as in the case of public goods (see note 93), due to the literature's employing combinations of regulation and income tax adjustments under which greater redistribution occurs. See Kaplow (2004, 2006c).

formulation of commodity taxes the benchmark rather than no differentiation, any deviation from the benchmark will, as before, be inefficient. Specifically, if one removes any deviations—that is, reforms commodity taxes to equal first-best Pigouvian taxes—and adjusts income taxes to offset distributive effects, a Pareto improvement will result. See Kaplow (2006c). The proof essentially tracks that presented in subsection 4.2 for commodity taxes. The main difference is that, as individuals adjust their consumption, this now changes the level of externalities and also changes the revenue from commodity taxes and subsidies (which are not moved to zero, but instead to first-best Pigouvian levels). Note, however, that the net revenue produced by the latter effect precisely equals the revenue necessary to compensate individuals for the former effect, through the income tax adjustment that holds everyone's utility constant. Therefore, just as in the case without externalities, optimal commodity taxes are those that otherwise would be efficient on first-best grounds.

As explored in subsection 8.2, the analysis of Pigouvian taxes and subsidies illuminates regulation much more broadly; hence, it would seem, as a benchmark, that regulations should be set with regard to the efficiency with which they mitigate externalities, independent of distributive effects and labor supply distortion.[104] One particular regulatory instrument is the use of legal rules, such as rules of tort liability. Indeed, as explored in subsection 8.2, a legal rule of strict liability is much like a Pigouvian tax (at least with regard to the injurer). In the analysis of legal rules, it has long been controversial whether adjustments should be made on account of distributive effects. Shavell (1981), in one of the first papers using the method of offsetting tax adjustments, shows that it is inefficient to deviate from otherwise efficient legal rules on account of distribution (which is held constant by the offsetting income tax adjustment). See also Kaplow and Shavell (1994).

9. Additional dynamic issues

Dynamic considerations were first introduced in extending the basic optimal income and commodity tax model to capital taxation in subsection 5.1. In this section, a number of further dimensions and complications that arise in a dynamic setting are examined.

9.1. Inflation[105]

Inflation complicates measurement of the tax base under a standard income tax because changes in wealth over time are subject to tax and the most readily available measures of such changes are denominated in nominal prices at different points in time. Note

[104] The claim regarding distributive effects is first advanced by Zeckhauser (1981). Distribution and labor supply distortion are addressed more fully in Kaplow (1996c, 2004, 2006c).

[105] See generally Aaron (1976), Feldstein (1983), and Halperin and Steuerle (1988).

that an implication of the source of the inflation problem is that a labor income tax and, likewise, a cash-flow consumption tax do not require significant modifications to adjust for inflation because the tax bases are appropriately measured in terms of current prices. (In nonlinear systems, just as under a nonlinear standard income tax, one would still need to index the bracket levels to avoid "bracket creep," wherein the real level of unindexed bracket boundaries declines over time as a consequence of inflation.)

One sort of inflation adjustment necessary in a standard income tax is indexation of basis. For example, if an asset was purchased in period 1 for $100 and is sold in period 2 for $125, the nominal gain of $25 will overstate the real gain if there is, say, 10% inflation. Basis adjustment involves restating the period-1 purchase price of $100 as $110 in period 2, so that the resulting taxable gain of $15 properly measures the real gain. Similar (ongoing) basis adjustments are necessary to preserve the real value of depreciation and to properly measure changes in inventories.

Another type of adjustment involves interest and related returns (which can alternatively be accomplished in an accrual system through basis adjustments to debt and related instruments). Suppose, for example, that inflation is 10%, the real interest rate is 2%, and the nominal interest rate is 12%. A 50% income tax applied to nominal interest would produce an after-tax nominal return of 6%, which would be a real after-tax return of −4%. Failure to index, therefore, results in an effective tax rate of 300% on the real return of 2%. Relatedly, borrowers, who deduct nominal interest payments, receive tax bonuses to a similar extent.[106]

Most standard income tax systems fail to index basis or account for the inflationary component of interest and related returns. They may do so indirectly, such as by providing a capital gains preference and more rapid depreciation, although such means often are not adjusted as inflation changes over time, they do not cover all pertinent dimensions, notably, the problem of interest, and they inevitably are more complete for some assets than for others.[107] As a consequence, when inflation is nontrivial, standard income taxes in practice deviate substantially from their idealized form and significant distortions can arise.

Indexing is more commonly employed by societies experiencing hyperinflation. In such settings, additional dimensions, in principle relevant even with low inflation, become significant, such as the difference in time between the withholding of taxes and payment to the tax authority and the fact that earnings and expenditures occur at different times within the standard accounting period, which is typically one year in length.

[106] Symmetry in theory can lead to asymmetric effects in practice because of incentives for high marginal tax rate actors to borrow from tax-exempt entities.

[107] LIFO accounting approximates the result of inflation adjustment as long as firms' inventories are not declining. Observe further that, as outlined in note 46, much capital income is effectively exempt under the existing income tax, so the inflation problem is limited to the remainder. However, this suggests that there may be even greater differentials in the taxation of capital than seem apparent, with some capital exempt and some of the rest effectively taxed at higher than stated rates.

9.2. Risk-bearing[108]

9.2.1. Uncertain labor income

Uncertainty in labor income provides a supplemental, efficiency-based justification for some redistributive income taxation.[109] Even if everyone is identical ex ante, individuals would prefer to have income transferred from high- to low-earning states, which a redistributive income tax does.[110] Given that existing income taxes and optimal schemes examined in the literature involve high tax rates even without uncertainty, the optimal adjustment for uncertainty may not be that large, both because the preexisting distortion is significant and because quite substantial implicit insurance would already be made available.[111]

The general problem of optimal income taxation in the presence of uncertainty has not been the subject of extensive study. Mirrlees's (1990) preliminary analysis (examining a linear income tax where the degree of variation in skill and the extent of uncertainty are assumed to be small) suggests that, taking as given the total variation in observed income, greater income uncertainty most plausibly favors a lower tax rate. For given aggregate variation, higher income uncertainty implies less variation in skill, and it turns out that skill variation is more powerful than income uncertainty in leading to a higher optimal tax rate.[112]

There are two caveats regarding the use of the income tax as insurance against uncertain labor income. First, to the extent that income uncertainty involves systematic risk, the government is not able to solve the problem: Its resulting budget uncertainty must be addressed, for example, by raising taxes or reducing spending if the resolution of uncertainty is adverse.[113] Second, the standard analysis ignores private insurance, the possibility of which renders government insurance through taxation not only unnecessary but also inefficient. Specifically, if the main inhibitor of private income insurance is

[108] Technically, one can analyze risk as resolved instantaneously, although as a practical matter the issues examined herein typically arise in a dynamic setting.

[109] See, for example, Eaton and Rosen (1980a, 1980b, 1980c), Tuomala (1990), and Varian (1980). Subsequent work includes Strawczynski (1998) and Low and Maldoom (2004).

[110] Varian's (1980) simulations suggest that, in light of moral hazard (i.e., the labor-leisure distortion), the optimal level of insurance may be only a few percent, whereas Strawczynski's (1998) and Low and Maldoom's (2004) simulations suggest that high marginal tax rates may be optimal.

[111] An important special case of income uncertainty involves the possibility that one will become disabled and thus unable to earn income in the future. See, for example, Diamond and Mirrlees (1986). Unemployment insurance is also pertinent.

[112] Tuomala (1990, section 9.3) finds that uncertainty in wages at the time effort is chosen (capturing such decisions as investment in human capital and occupational choice) favors rising marginal tax rates, the intuition being that high realizations are substantially attributable to luck, so the disincentive effect of taxing them more heavily is modest. His results are surprising in that the extent to which this is true in his simulations actually increases as uncertainty falls, which seems to contradict both the given intuition and the results of his simulations elsewhere in his book, which generally display falling marginal rates when there is no uncertainty.

[113] Compare Bulow and Summers (1984) and Gordon (1985).

moral hazard, as much of the literature asserts, the government cannot combat it either; indeed, that is why there is a labor-leisure distortion from income taxation. However, if adverse selection or other imperfections impede private insurance, then government insurance such as through taxation may be optimal. It should be kept in mind, however, that private and direct government insurance does exist for some important sources of uncertainty in labor income, notably for disability and temporary unemployment, and various other means allow individuals to mitigate income uncertainty to some extent.[114]

9.2.2. *Uncertain capital income*[115]

Uncertainty is central to the understanding of capital income taxation because a substantial portion of the return to capital, notably regarding equity investments, consists of a risk premium. Domar and Musgrave's (1944) seminal paper offers a partial equilibrium analysis in a model with two assets, one riskless and the other risky. Their basic insight is that taxes on the risky component of returns have no effect on individuals, who will simply gross up their investments in the risky asset to offset the effect of the tax.

To verify this result, suppose that, in a world in which only the riskless return, r, is taxed at the rate t_r, an investor would invest X in a risky asset that pays R_i (possibly less than X) in state of the world i. (For simplicity, ignore the rest of the investor's portfolio, which is held constant and which is unaffected by the reform to be considered.) Then, after risk is realized and tax is paid, the investor will have $R_i - t_r r X$, the gross return minus the tax on the riskless return to the initial investment.

Now assume instead that all investment returns, including the risky component, are to be taxed at the rate t_r (in a fully symmetrical manner, allowing for complete deductibility of losses). The investor can offset the effect of this supplemental tax on risk by increasing his investment in the risky asset from X to $X/(1 - t_r)$, borrowing the additional funds, $t_r X/(1 - t_r)$, at the riskless rate r. After risk is realized, the loan (with interest) is repaid, and tax is paid, the investor will have

$$
\begin{aligned}
&\frac{R_i}{1-t_r} - \frac{t_r X}{1-t_r}(1+r) - t_r\left[\frac{R_i - X}{1-t_r} - \frac{t_r X}{1-t_r} r\right]\\
&\quad = (1-t_r)\frac{R_i}{1-t_r} - \frac{t_r X}{1-t_r}(1 + r - 1 - t_r r)\\
&\quad = R_i - \frac{t_r X}{1-t_r} r(1-t_r) = R_i - t_r r X. \qquad (9.1)
\end{aligned}
$$

In the first line of expression (9.1), the first term is the gross return on the investment of $X/(1 - t_r)$, the second term is the repayment of the loan, $t_r X/(1 - t_r)$, with interest, and the third term is the tax owed, the tax at rate t_r being levied now on the gross return net of the investment, with a deduction allowed for the interest payment on the loan. As can

[114] See Cochrane (1991) and Mace (1991).

[115] On taxation, risk, and household portfolio behavior more generally, see Poterba (2002).

be seen, in every state (for any i), the investor's net return under this regime—having adjusted the initial level of investment—is $R_i - t_r r X$, which is identical to the net return under the initial regime that taxes only the riskless return to investment.

This basic model has been extended in various ways that, among other things, take account of general equilibrium effects in asset markets and whether the government's budget is in balance in different states of the world. See Bulow and Summers (1984), Gordon (1985), and Kaplow (1994). An important implication of this model is that capital taxation (see subsection 5.1) is less important than it may appear to be because a tax on all returns is, accounting for portfolio adjustments, equivalent to a tax on only the riskless return, which is but a portion of the total return to capital. In addition, permutations of the foregoing analysis imply that various equivalences among taxes (for example, between a labor income tax and a consumption tax) that hold in a static model and in a dynamic world with certainty extend to the case of uncertain capital income. See Kaplow (1994).

9.2.3. Other losses

Aside from uncertainty that may affect the return to labor effort and investment, individuals may suffer losses directly, notably in cases of illness, requiring possibly significant medical expenditures, and casualties, such as if one's home is destroyed by fire. A common view, reflected to an extent in some income tax systems through deductions, is that individuals' effective income is lower on account of such losses and hence their taxable income should be reduced accordingly.

There are two shortcomings in this logic. First, the point has not been developed properly in an optimal income tax model. Even if the marginal utility of consumption of an individual who, say, earns \$50,000 and suffers a loss of \$10,000 is the same as that of someone who earns \$40,000 and suffers no loss, their ability level and thus various aspects of the optimization are different, so further analysis of the optimal treatment is required. Second, a complete analysis must take account of the possibility that individuals could insure or take various precautions ex ante. See Kaplow (1992). Providing a deduction for losses is tantamount to providing free insurance for a fraction of losses equal to an individual's marginal tax rate. Such implicit insurance produces moral hazard, which distorts private insurance decisions. If moral hazard is the only market imperfection (or, even more so, if private insurance can combat moral hazard, such as where some precautions are observable), insurance through the tax system is inefficient. Furthermore, if the tax deduction is limited to uninsured losses, as in the United States income tax, private insurance is inefficiently discouraged.

9.3. Transitions and capital levies[116]

Many contemplated fundamental tax reforms involve transitions that entail what is tantamount to a one-time capital levy (or grant). One-time capital levies are traditionally

[116] See Kaplow (2007a) on these and other transition issues; see also Shaviro (2000).

viewed as an ideal sort of tax. Such a levy is imposed only on preexisting capital and, being one-time (with a presumed credible commitment never to be repeated), is nondistortionary.

There are two major problems with arguments in favor of a capital levy. First is the familiar point that the future promise may not be credible. The prospect of capital levies is a serious fear in many developing economies, which discourages foreign investment and induces residents to send capital outside the country. Any government that actually imposes a capital levy would not expect to be trusted anytime soon. That most countries refrain from such policies reflects some mix of constitutional limitations, strong norms, and the fear that future capital flight would be more costly than any short-run, even if substantial, gain.[117]

Second, there is a conceptual problem with the purported idyllic nature of a capital levy in developed economies that have an income tax.[118] Specifically, one could in principle raise sums distortion-free by reducing the grant component g of the income tax, even making it negative—i.e., a uniform lump-sum levy. The primary deterrent to using this approach is not inefficiency but dislike of the distributive consequences. If a capital levy is a mere substitute for lowering g, and if g is already set optimally, then there is no benefit to a capital levy.

Further reflection suggests that a capital levy may nevertheless be welfare-increasing (if the one-time feature were realistic) because, at any given point in time, it is likely that ownership of the existing capital stock is distributed in a way that positively correlates with income and underlying earning ability, for it constitutes the result of prior accumulations that ultimately derive from labor income. (The correlation will be highly imperfect due to differences in life-cycle stage, preferences, and other factors.) Thus, a capital levy may be distributively appealing whereas reducing g would not be. Observe that this version of the argument is closely analogous to the notion that it would be ideal to impose future individualized lump-sum taxes based on ability as inferred from prior earnings or investments in human capital: The taxes would be nondistortionary in the future, and as long as this regime was not anticipated ex ante, society would not have suffered from prior distortions either.[119] If such a one-time imposition were possible, it would be optimal to fashion the redistributive tax as a function of revealed labor effort rather than imposing a uniform capital levy.

In sum, one-time capital levies may seem attractive, depending on the available alternatives, but are generally regarded as infeasible, if not dangerous even to contemplate actively. It is interesting, therefore, that many fundamental tax reforms can involve what is tantamount to a capital levy. Most analyzed is the transition from an income tax to a consumption tax (or simply the introduction of or raising the rates in a consumption tax): Unless there is transition relief for pre-enactment accumulations, the effect is to

[117] For the classic statement of the dynamic commitment problem, which mentions capital levies as an illustration, see Kydland and Prescott (1977).

[118] Compare the discussion of Ramsey taxation in subsection 4.4.

[119] For a theoretical exploration of taxation based upon prior economic choices, see Roberts (1984).

reduce the purchasing power of preexisting capital.[120] Thus, although a wage tax and a consumption tax may be equivalent in steady-state, simulations of the transition to a consumption tax show greater efficiency gains than result from transition to a wage tax because the former contains a significant capital levy whereas the latter does not.[121] Of course, if the transition were anticipated, say during years or decades of preceding debate, the implicit levy would not be unanticipated and distortion would result.

Similar questions arise in other settings. For example, analyses of the efficiency of capital taxation more generally often envision a world in which there is a preexisting capital stock without inquiring as to its origin—notably, as a consequence of prior earnings or of inheritance, which itself is the product of a donor's prior earnings, on which see subsection 5.2. In such a model, the intertemporal inefficiency of taxing capital in a simple setting (see subsection 5.1) will be counterbalanced by the advantage of the capital levy, and a dynamic analysis will accordingly suggest the seeming optimality of high capital taxes initially but no capital taxation in the long-run steady state.[122] All of this, of course, assumes that the enactment of the initially high capital tax is unanticipated and that the subsequent promise to eliminate capital taxation is credible.

Likewise, negative capital levies—windfalls—can arise if, for example, corporate taxation is reduced or eliminated. Thus, a key point concerning the desirability and appropriate form of integration of the corporate income tax (see subsection 5.1.3) concerns the fact that existing corporate equity would thereby be freed of future tax liability. This revenue loss is ordinarily seen as unaccompanied by any corresponding efficiency gain, although in a setting in which the possibility of integration may long be anticipated, an understanding that old equity would benefit would tend to have the effect of reducing pre-enactment distortion.[123]

9.4. Capital gains

The attempt to employ an accrual income tax (taxing returns to both labor and capital) is plagued by the problem of capital gains. Proper taxation requires that gains be taxed as they accrue or, in the alternative, following Vickrey (1939), that interest be charged on taxes that are deferred until realization (ordinarily, the sale of the asset). Ongoing accrual taxation (referred to as "mark to market") is feasible for many publicly traded securities but not for some other assets, and selective application of accrual taxation would be distortionary.[124] As a consequence, most assets are taxed on a realization

[120] See, for example, Bradford (1996a, 1996b), Kaplow (2007a), Sarkar and Zodrow (1993), and Shaviro (2000).

[121] See, for example, Auerbach and Kotlikoff (1987) and Sarkar and Zodrow (1993).

[122] See, for example, the results reported in note 41.

[123] Expansions and contractions of social security benefits can be analyzed in an analogous manner: If taxpayers perceive a significant tax-benefit linkage and if (but only if) benefit changes are anticipated, this prospect should affect pre-enactment labor supply.

[124] Possible lack of liquidity is another reason sometimes given for disfavoring accrual taxation, although the problem is unlikely to be significant for publicly traded assets for which mark to market is feasible.

basis, resulting in mismeasurement of income (which may cause interasset distortion since the benefits of deferral vary across types of assets) and distortion of investors' portfolios on account of the lock-in effect, wherein taxpayers with accrued gains have an incentive to defer sale in order to further defer taxation.[125] Additionally, the realization requirement induces a variety of financial manipulations designed to take advantage of its benefits.[126]

An ingenious scheme, initially due to Auerbach (1991) and subsequently extended by Bradford (1995) and Auerbach and Bradford (2004), solves the problem using a sort of realization-based taxation that, in its most basic form, taxes investors as if the value of the asset at the time of sale was produced by a hypothetical investment at the time of purchase that grew at the riskless rate of interest. Two desirable features of this approach are that relative risk-adjusted asset returns are unaffected, just as under an idealized accrual tax, and lock-in is avoided. An apparent shortcoming is that actual tax payments diverge from what one would think should be due in particular states of the world. For example, if there is a huge gain, the investor's tax based on the imputed riskless return is far less than what would be due under an actual accrual tax, and if there is a loss, the investor still pays positive tax based on the imputed riskless return. However, the central virtues of the tax scheme are unaffected by these apparent anomalies. Moreover, when one takes into account individuals' portfolio adjustments (compare subsection 9.2.2), investors' actual net positions in each state of the world are the same as they would be under an ideal accrual tax. See Kaplow (1994).

9.5. *Human capital*

Human capital constitutes a substantial majority of all capital. See, for example, Davies and Whalley (1991) and Jorgenson and Fraumeni (1989). Furthermore, the returns to human capital—wages—are the direct or indirect source of most tax revenue. Accordingly, the relationship between human capital and taxation deserves significant study. Nevertheless, the subject has received far less attention than has the taxation of physical and financial capital. To illuminate the matter, it is useful to compare the tax treatment of human capital under an accrual income tax, which purports to tax capital income, with standard treatments of physical and financial capital, following Kaplow (1996a).[127]

[125] That the United States also provides a tax-free step-up of basis at death adds to the lock-in problem. Note that there is an important exception to the realization approach for business assets, namely, the provision of depreciation, amortization, and other deductions for business expenditures on capital assets that predictably decline in value over time. (In the absence of such provision, taxpayers would have ongoing incentives to sell used plant and equipment, to realize losses.) To the extent allowed deductions depart from economic depreciation (on which see Samuelson (1964)), interasset distortions result. One case involves intangibles, such as investments in advertising and R&D, which are permitted to be expensed.

[126] One of the simplest is the selective sale of assets with losses so that they can be deducted against other income, while continuing to hold assets with unrealized gains. See Stiglitz (1985a). Capital loss limitations are employed to counter this tactic.

[127] For a complementary treatment of some of these issues, see Andrews and Bradford (1988, appendix).

An accrual income tax, which taxes both labor income and returns to capital, treats human capital essentially on a realization basis by taxing wages while ignoring changes in the value of an individual's stock of human capital. This is most apparent in an individual's last year of work: The year's wages are taxed, but the individual's stock of human capital will have fallen during the year by approximately the full amount of these wages (the present value at the outset equaling the year's wages with a slight time value discount) while no offsetting depreciation deduction is allowed, such as would be the case under pure accrual income taxation. Under an idealized accrual system, the receipt of human capital (at birth) would be subject to tax, and each year's earnings would be partially offset by depreciation deductions that would be growing over time. Further taxation (including negative taxation for falls in human capital) would arise as uncertainty was resolved, just as with a physical or financial asset that produced a similar (yet equally uncertain) pattern of future cash flows.

To highlight the difference between accrual taxation of human capital and the realization-based approach implicit in taxing wages instead, with no adjustments for changes in the value of human capital, one can ask, just as in subsection 9.4's discussion of capital gains, what taxation at realization would serve as a proxy for accrual taxation. As a crude approximation, by analogy to Auerbach's (1991) scheme, any wage earnings would be subject to greater tax the longer the holding period, i.e., the older the individual earning the wages. Thus, pure accrual taxation would be similar to applying to wage earnings a multiplier that grew over individuals' lifetimes.

Given that actual taxation of human capital is realization based (with no multiplier) and thus closer to the treatment appropriate under a consumption tax, we can ask what are the efficiency implications of this attribute. First, to the extent that consumption taxation is more efficient by reducing intertemporal distortion (see subsection 5.1), this form of wage taxation may seem desirable. The discrepancy with the treatment of other capital suggests that individuals would prefer to fund future consumption by deferring labor supply to later years rather than by earning more now and engaging in conventional savings. Because later years' wages are not taxed at a higher rate reflecting the implicit deferral whereas conventional savings are subject to tax on appreciation under an accrual income tax (and to a lesser extent under existing, standard income tax systems), there does exist a preference for saving through deferral of labor supply. This secondary distortion of labor supply, in the timing of earnings, serves to partially offset the distortion of intertemporal consumption choices under an accrual income tax.

Under standard forms of capital taxation, such as in an accrual income tax, investment in human capital would appear to be favored over investment in other capital because human capital is only taxed upon realization. As noted, however, there are no depreciation deductions for human capital and thus for incremental investments therein. Some investments in human capital, notably forgone earnings, are implicitly expensed since the forgone imputed income is never taxed, which is more favorable than depreciation. See Boskin (1977) and Heckman (1976). However, if such implicit deductions are taken in years in which marginal rates are low (due to rate graduation), the result could be less generous. See Nerlove et al. (1993). Many direct investment expenditures, such as on education, are never deductible (although there are also substantial public subsidies). Additionally, to the extent that the uncompensated labor supply elasticity is positive (negative), the reduction (increase) in labor effort due to labor income taxation will reduce (increase) the value of investments in human capital. Further influences on the return to human capital investment may arise on account of general equilibrium effects on wages of the sort noted in subsection 3.5.2. The net effect of these (and other) factors is not entirely clear. Nevertheless, Trostel (1993) estimates that income taxes do significantly discourage investment in human capital, the primary channels being a consequence of the taxation of labor income, implying that a consumption tax would have a similar effect. It does not follow, however, that all such effects involve inefficiencies: Taking as given that individuals, say, will work less on account of labor income taxation, it is efficient for them to invest less in their human capital to that extent.[128]

Human capital is also treated qualitatively differently from other capital under transfer tax systems. See Kaplow (2001). Many countries subject sizeable gifts and estates, usually transfers to the next generation, to high marginal tax rates. However, transfers of human capital are largely exempt. The primary constituents—genetic endowment, environment (parental involvement and schooling), opportunities, and contacts (see Taubman (1996))—are not ordinarily considered as conceivable components of a transfer tax base. Direct expenditures, such as on private education or on more expensive housing that gives access to superior public schools, are typically exempt as well. As noted in subsection 5.2, however, the negative externality of gifts due to the income effect may be reversed in the case of gifts of human capital (and the standard positive externality to the donee remains), so preferential treatment of transfers that contribute to human capital may be optimal. Of course, there may be some inefficient discrimination against other transfers having similar effects, such as those that relax liquidity constraints and thereby facilitate entrepreneurship.

9.6. *Lifetime horizon*

An important simplification implicit in standard, static optimal income tax analysis is that individuals' lives, both their labor effort and consumption, are collapsed into a single period. However, because individuals' earnings vary over the life cycle both systematically and on account of uncertainty and because individuals borrow and save to allocate (generally, to smooth) consumption over the life cycle, much of relevance is omitted as a consequence. In particular, with income taxes that are assessed on an annual basis, the mismatch between current and lifetime distributive effects is significant. Many of the current "poor" consist of young or old individuals with low current but high lifetime income or individuals of any age who may be having a bad year, whereas

[128] There may be externalities to investment in human capital, and general equilibrium effects also influence welfare, so such effects of taxation on human capital may still be of social consequence. See, for example, Hamilton (1987), Heckman, Lochner, and Taber (1999), and Jacobs (2005).

some of the "rich" may be moderate-lifetime-income individuals having a good year. With hump-shaped earnings profiles, even middle-lifetime-income individuals with a certain wage stream will have below-average incomes when young and when old and above-average incomes when in their peak earning years.[129] See generally Fullerton and Rogers (1993).

If individuals engaged in no borrowing and lending (and there were no consumer durables and no other sources of utility interdependence across periods), the optimal income tax problem would be largely unaffected: In each period, individuals would have a given wage and the social welfare optimization could proceed period by period. If, however, there is significant borrowing and lending (directly or indirectly), as clearly occurs, the problem changes. The optimum would involve age-dependent tax schedules, and the optimization would need to take into account individuals' consumption-smoothing behavior, including possible limits to theoretically ideal smoothing due to liquidity constraints and myopia.[130]

Short of a complete analysis, it is widely suggested, following Vickrey (1939), that some sort of averaging scheme would be appropriate. Lifetime averaging may seem ideal, and following Vickrey's proposal is less complex than it might seem, although changes in family unit membership over the life cycle do make the problem significantly more challenging. See Liebman (2003) for estimates of the effects of averaging over various periods under the current United States tax and transfer system. Liebman also emphasizes the need to relate income averaging explicitly to social welfare, noting that averaging can produce not only distributive benefits but also efficiency gains, on account of equalizing marginal tax rates across individuals and over time.

Two caveats regarding the importance of averaging should be noted. First, if the redistributive tax is based on consumption rather than labor income or total income, the potential mischief caused by focusing on annual information is greatly mitigated on account of consumption smoothing. In the simplest case (including a world with complete earnings certainty and perfect annuity markets), individuals would consume evenly over their lifetimes, so no adjustments would be required. Second, even under an income tax, variable earnings over the life cycle would not have the ordinarily posited effects if the

[129] For similar reasons, the incidence of many forms of taxation is more nearly proportional (rather than progressive or regressive) if the basis of comparison is lifetime rather than annual income (or if it is consumption, which over the life cycle is smoother than income). See Fullerton and Rogers (1993). There are also more subtle effects. For example, median-lifetime-income individuals have a later earnings peak than high- or low-income individuals and hence save less when smoothing their consumption. As a consequence, Fullerton and Rogers find that the lifetime incidence of taxes on capital tends to be U-shaped, falling more heavily on the rich and the poor than on the middle class.

[130] The social security system (see subsection 5.3) is age-dependent, but primarily as a means of forced savings. It is unclear the extent to which its features, when combined with the annual income tax, are in accord with what would be optimal. On lifetime income, income taxation, and social security, see Diamond (2003).

tax schedule were linear.[131] Only with graduated rates are tax burdens higher as a consequence of uneven earnings. As explored in section 3, however, it is not apparent the extent to which highly nonlinear marginal tax rates are optimal.

9.7. Budget deficits and intergenerational redistribution

Another common assumption in much analysis of taxation is that budget balance is required. In a dynamic setting, this requirement need not be met in any particular accounting period but must hold in the long run; deficits today must ultimately be financed, even if by paying interest in perpetuity. The time pattern of deficits is not, however, a matter of indifference, even abstracting from macroeconomic effects.

First, Barro (1979) shows that it is advantageous to adjust short-run deficits so as to maintain constant tax rates (in expectation). Because the marginal distortionary cost of taxation rises with marginal tax rates, it is better to raise revenue over time with tax rates that are as nearly constant as possible rather than with more highly variable tax rates, just as in certain basic static settings it is best to tax different activities uniformly. Accordingly, whether a deficit makes sense today depends on whether expenditures are temporarily high (for example, on account of a war), on projections for future growth of the tax base, and on other factors bearing on whether paying off the deficit will require higher tax rates in the future.

Second, different timing of deficits has distributive effects and, in particular, is one potential source of intergenerational redistribution. For this reason, Auerbach, Gokhale, and Kotlikoff (1991) and others have proposed to track such effects using generational accounting, wherein the aggregate effect of government tax and transfer policy (and, in some work, certain government expenditures, such as on education) over time is imputed to different generations, making the assumption that deficits or surpluses of the past and present will ultimately be borne by future generations. See Kotlikoff (2002). Social security (see subsection 5.3), when funded largely on a pay-as-you-go basis, obviously has important intergenerational effects. Furthermore, uncompensated transitions that may involve large capital levies (see subsection 9.3) can likewise have distributive effects, such as by heavily taxing existing capital, owned disproportionately by older living generations, which allows reduction of the national debt that otherwise would ultimately be paid off by younger or future generations. An important complication, related to debates about Ricardian equivalence, see Barro (1974), concerns the extent to which private intergenerational transfers (variations in the level of gifts and bequests) will adjust to offset changes in public intergenerational transfers.

[131] There still is an important effect in an income tax that includes both labor and capital income because, as noted in subsection 9.5, earnings in earlier years of equal present value are disfavored by the tax on the return to savings.

10. Unit of taxation

10.1. Framework

Treatment under income taxes and transfer programs depends on the type of family unit, notably, whether there is a married couple or a single adult and whether and how many children are present. What treatment is proper has proved quite controversial, with actual practice varying substantially among programs, across jurisdictions, and over time. This section considers work that seeks to relate issues involving the unit of taxation to the explicit welfare-maximizing approach that serves as the basis for optimal income tax analysis when there is only one type of taxable unit.

The framework and analysis outlined here and developed in subsections 10.2 to 10.5 follows Kaplow (1996b). It simplifies the problem by holding labor supply and family composition fixed (on which see subsection 10.6), thereby focusing on the question of how the relative treatment of different types of family units affects social welfare on account of distributive effects.[132] It is helpful to consider a simple model with two family units, a single individual and a two-member family: For some applications, the two members may be thought of as two adults, and for others as a representative parent and a representative child. With labor supply and thus total income fixed, it is assumed that the sole policy choice is an allocation between the two units. Allocations within the two-person family are assumed to be determined in some internal fashion that the government is unable to observe or control.[133]

To complete the description of the model, it is necessary to state the pertinent utility functions, the mechanism by which sharing within the two-person unit is determined (see, for example, Becker (1991)), and the SWF.[134] For the latter, analysis will consider the utilitarian case; the implication of a strictly concave SWF will be apparent below because most results depend on the concavity of individuals' utility and a more concave SWF will usually have implications similar to those of more concave utility functions.

As a benchmark, consider the case in which all individuals—the single individual and both family members—have the same utility functions, there is equal sharing in

[132] A further aspect of non-fixed family composition is that individuals spend different parts of their lives in different types of family units, so optimization over the life cycle raises additional complications in determining optimal taxation of different family units.

[133] Allocations can be influenced by differential commodity taxation (assuming that spouses or children have different demands) and by providing transfers in-kind.

[134] Two features are noteworthy. First, each individual is the unit of analysis in the SWF, rather than referring to a family utility function (although one could define the latter as the sum of the utilities of the family members). Second, this method does not involve defining an equivalence scale and considering, say, how to equalize equivalent income. See, for example, Deaton and Muellbauer (1986) and Gronau (1988). As will become clear, this alternative approach would not be tantamount to maximizing any standard SWF, except by coincidence. For criticism of the use of equivalence scales for welfare analysis, see, for example, Pollak and Wales (1979).

the (two-person) family, there are no economies of scale, and there are no interdependencies in utility functions. As with Edgeworth's (1897) egalitarian result, the optimum here obviously involves giving the single individual half the consumption as the family because this allocation equalizes everyone's marginal utility of consumption. Note that everyone's total utility is the same as well.

10.2. Intrafamily sharing

The optimal treatment of different family units may well depend on how sharing operates within multi-person families.[135] Indeed, in debates about whether married couples should be taxed as a unit (as in the United States) or as if they were two single individuals (as in many other countries), it is often suggested that the former treatment is more appropriate to the extent that sharing is equal. (A primary reason the difference matters is due to graduated rates, which make separate treatment less advantageous;[136] note, however, that the analysis in subsection 3.4 hardly indicates that graduated rates are optimal.) Such commonly held views, however, consider neither how the resulting differences in tax burdens will actually be shared nor how the results relate to maximization of a standard SWF.

Insight into the issue can be gleaned by beginning with the above-described benchmark case and assuming now that the two-person family shares resources unequally in some stipulated ratio, with ϕ being member 1's share and $1 - \phi$ being member 2's. Without loss of generality, consider the case in which $\phi > 1/2$. Furthermore, because we are abstracting from labor supply and considering identical utility functions, we can let u denote each of the individual's utilities as a function of post-redistribution consumption. It is also convenient to let c denote the total consumption of the two-person family and ωc that of the single individual.[137] This allows the simple interpretation that, when ω equals (or is greater or less than) $1/2$, the single individual receives precisely (or more or less than) a pro rata share of total resources.

Social welfare is equal to $u(\phi c) + u((1 - \phi)c) + u(\omega c)$. Differentiating with respect to c (keeping in mind that ω is implicitly a function of c) yields the following condition for the utilitarian optimum:

$$\phi u'(\phi c) + (1 - \phi)u'\big((1 - \phi)c\big) = u'(\omega c). \tag{10.1}$$

[135] See Apps and Rees (1988). Note also that when family members' earnings differ from their consumption allocations, gifts or exchange are implicitly involved. Hence, the question of taxation of different family units must, in principle, be related to the question of transfer (gift and estate) taxation, see subsection 5.2, especially in light of the fact that a substantial portion of all voluntary transfers are to other family members. This relationship is illustrated in subsection 10.4 on altruism.

[136] To illustrate, suppose that for single individuals there is a 20% marginal tax rate on income below $50,000 and a 40% rate above, and that married couples face the same rates with the upper tax bracket beginning at $100,000. If one individual earns more than $50,000 and another earns less than $50,000, marriage will reduce their combined tax payments because part of the former's income that would have been taxed at 40% will now be taxed at 20%.

[137] That is, letting C denote total resources, $\omega = (C - c)/c$.

This first-order condition, as one would expect, sets the sum of the two family members' marginal utilities from an additional dollar allocated to the two-member family (taking account of how much of the dollar each member actually will consume) equal to the marginal utility of an additional dollar allocated to the single individual. The first term on the left indicates the benefit from the fraction ϕ of a dollar allocated to the family going to the first member, whose marginal utility reflects the consumption share ϕ. The second term likewise indicates the benefit from $1 - \phi$ of a dollar going to the second member. The right side is the single individual's marginal utility of a dollar.

It is indeterminate whether ω exceeds, equals, or is less than $1/2$. The first term on the left side of (10.1) is the effect of allocating additional funds to the family through its effect on the individual who takes the greater share. Since that person receives more than a pro rata share ($\phi > 1/2$), this term receives more weight; however, for the same reason, that person's marginal utility will be lower, on account of the diminishing marginal utility of consumption, giving the term less weight. For the second person, these effects reverse. Which effect is larger depends on the curvature of the utility function. Taking the constant-relative-risk-aversion case, if the coefficient of risk aversion exceeds (is less than) 1, it can be shown that the allocation should be more (less) favorable to the family—ω should be less (greater) than $1/2$. The intuition is that, when the curvature is greater, it matters more that some of the additional resources to the family are enjoyed by the less advantaged member because that member's higher marginal utility of consumption more than counterbalances the fact that a majority of the added resources go to the more advantaged member.

This result assumes that, whatever is each family member's share, it is constant rather than a function of consumption. If the share of the more-advantaged member rises (falls) with consumption, treatment of the (two-member) family should be less (more) generous than implied by expression (10.1). This is because the marginal dollar more (less) favors the advantaged member than does the average dollar, and the average dollar—really the total—determines individuals' marginal utilities of consumption. In sum, to determine what treatment is optimal on account of unequal sharing requires specifying (and justifying) a particular model of how intrafamily sharing operates. See, for example, subsection 10.4 and the surveys by Behrman (1997) and Bergstrom (1997).

10.3. Economies of scale

Economies of scale are thought to be the main justification for applying a higher tax schedule to married couples than to (two) single individuals, such as is done in the United States. To assess this conventional wisdom, return to the benchmark case (with equal sharing) and suppose that a dollar of consumption is worth more to the two-member family. Specifically, assume that each of the two family members' utility is given by $u(\beta(c)/2)$, where $\beta(c) > c$. For concreteness, consider the linear case: $\beta(c) = \beta c$, with $\beta > 1$.

The first-order condition for welfare maximization is now

$$\beta u'(\beta c/2) = u'(\omega c). \tag{10.2}$$

Again, there are two competing effects. On the left side, the leading β indicates that the two-member family gets more per dollar than does the single individual, which favors more generosity toward the family ($\omega < 1/2$). But the argument of u' on the left side is $\beta c/2$ rather than $c/2$, a higher value, which reduces the marginal utility of consumption, favoring less generosity ($\omega > 1/2$). (This latter effect corresponds to the conventional wisdom on the subject.) Which effect is greater depends again on the curvature of the utility functions, although the results are opposite to those with unequal sharing. With constant-relative-risk-aversion utility functions and a risk-aversion coefficient above (below) 1, the latter (former) effect dominates, making less (more) generous treatment of the family optimal. The intuition is that, when curvature is high, the higher standard of living due to scale economies greatly diminishes the marginal payoff to additional consumption relative to the multiplier accounting for the family's greater efficiency of converting tangible resources into effective consumption.

10.4. Altruism

One specific case of interest is that in which each family member is altruistic toward the other. Specifically, suppose that the utility of each member of the two-person family is given by $u_i + \psi_i u_j$. Then the total utility entering into the SWF from the two-member family will be $(1 + \psi_2)u(\phi c) + (1 + \psi_1)u((1 - \phi)c)$. For preliminary insight into the effect of altruism, assume equal sharing ($\phi = 1/2$). Then, the first-order condition is

$$\left(1 + \frac{\psi_1 + \psi_2}{2}\right)u'(c/2) = u'(\omega c). \tag{10.3}$$

Clearly, the allocation should be more favorable to the two-member family ($\omega < 1/2$) to an extent that increases with the level of altruism. Here, the two-member family, as in the case of scale economies, more efficiently converts tangible resources into utility. However, given the stipulated manner of the conversion, there is no offsetting effect from a diminished marginal utility of consumption.

One suspects that stipulated equal sharing may be inconsistent with altruism, especially when altruism is asymmetric. Another simple possibility is that the shares are chosen to maximize family welfare. Compare Samuelson (1956). Making use of the first-order condition for the family's optimization problem, one of the equivalent ways to express the first-order condition for social welfare maximization is

$$(1 + \psi_2)u'(\phi c) = u'(\omega c). \tag{10.4}$$

(The parameter ψ_1 implicitly enters, symmetrically, through the determination of ϕ in the family's first-order condition.) It can be shown, as one would expect, that greater altruism favors more generosity to the family.

Another interesting case is that in which only one member (member 1) is altruistic and that member chooses ϕ to maximize personal welfare, as in Becker (1974). In this case, the optimal treatment will depend on the strength of altruism. For example, if $\psi_1 = 1$, which corresponds to giving equal weight to the other member's utility, equal sharing will result and the outcome of the preliminary case in this subsection, involving more generous treatment to the family, will govern. When $\psi_1 < 1$ but is still sufficiently large that some sharing occurs, the result is more complex. As in subsection 10.2, although few incremental resources may go to the less advantaged member, the marginal utility from those additional resources will be relatively high.

10.5. Children

If children were like adults (in terms of the model, meaning the same utility functions and so forth), then no additional analysis would be required. In the benchmark case of subsection 10.1, families' allotments would optimally involve each member getting the same resources as a single individual. Such treatment is substantially more generous than that provided by tax systems, which usually make relatively modest adjustments for children, especially for individuals above fairly moderate levels of income. (Indeed, in the United States, some adjustments are phased out at higher levels of income.) Existing treatment would be optimal if, for some reason, the welfare of children simply did not count as full constituents of social welfare, a rather implausible view.

Suppose that children differ because they require fewer resources to achieve a given level of utility. Specifically, assume that member 2's utility is given by $u((1-\phi)c/\zeta)$, where $\zeta < 1$. Member 1, taken as a representative adult, continues to have the same utility function and thus to achieve the same level of utility for a given level of consumption as does the single individual. First, consider the case in which the family's sharing equalizes the utility levels of the two family members. This requirement implies that $\phi = 1/(1+\zeta)$. The first-order condition for welfare maximization is

$$\frac{2}{1+\zeta}u'\left(\frac{c}{1+\zeta}\right) = u'(\omega c). \tag{10.5}$$

Because $\zeta < 1$, the leading component on the left side of expression (10.5) exceeds 1. Thus, it must be that $1/(1+\zeta) > \omega$. Furthermore, because $\phi = 1/(1+\zeta)$, we have $\phi > \omega$. This means that the optimal allocation is more generous to families than one that would equalize the total utility of each person—or, equivalently in this case, equalize the utility of member 1 and that of the single individual, each of whom has the same utility function. This result is due to the fact that member 2 is a more efficient generator of utility, and the only way to channel additional resources to member 2 (given the stipulated intrafamily sharing rule) is to benefit member 1 as well. (Again, the curvature of the utility function will also affect the optimum: The greater the curvature, the less is the efficiency effect relative to the diminishing marginal utility effect, so the less is the aforementioned preference for the family.)

Second, consider the case in which the family's allocation maximizes the sum of their utilities rather than equalizing the levels of their utilities. This allocation would be relatively more advantageous to member 2, the more efficient utility generator, and the share would be set at the level where the efficiency effect just equalled the diminishing marginal utility effect. The result is that the socially optimal allocation entails $\phi = \omega$, an allocation under which member 1 has the same resources and utility as the single individual. (The family equates the marginal utility of resources between the two members, and the social authority equates the marginal utility of the family, which equals the marginal utility of either member—and we can take member 1—to the marginal utility of the single individual. Since member 1 and the single individual have the same utility function, they receive the same allocation.) As in the prior case, however, the family's allocation is sufficient to allow member 2 to achieve a higher utility than that of the single individual (so average utility in the two-member family remains higher).

It would seem that optimal allocations to families with children are more generous than is typically provided. Whether and to what extent this is true depends on the actual form of the utility function. For example, if children are such efficient utility generators that they achieve high utilities with virtually no resources and are subject to rapidly diminishing marginal utility at that point, little adjustment for family size may be required.[138]

Furthermore, the above results suggest that family size adjustments should not be limited to lower-income families. It is often thought that extending or expanding adjustments to higher-income individuals (perhaps making them proportional to income) would be highly regressive. Such reasoning, however, is misleading in two respects. First, it ignores that the overall tax schedule can be adjusted simultaneously to avoid such an effect; thus, the relevant questions is, as between families of a given level of income, what should be their relative tax payments (or transfer receipts) as a function of the number of children. Second, any characterization of this sort presumes that one has already determined who is rich and poor and to what extent, but in setting adjustments for different family units, the point of the enterprise can be understood (loosely) as an implicit process of defining how rich or poor various family units are in the first instance.

10.6. Incentives

The foregoing analysis of optimal taxation of different family units takes labor supply and family composition as given. However, there are likely to be interactions between how different family units are taxed and each of these incentive margins.

First, consider work incentives. One way to view the optimal income tax problem when there are different family units (taken to be exogenous for the moment) makes use of the analysis in subsection 3.5.1 in which a different $T(wl)$ schedule is applied to each group (θ would now index family composition). The analysis in the preceding subsections can be understood as bearing on the form of individuals' utility functions for each group. In other respects, analysis would proceed as before.[139]

[138] Another complication is that all government policy, including for example free public education, must be taken into account in determining the extent of generosity in an existing system.

[139] Cremer, Dellis, and Pestieau (2003) examine a model with exogenous differences in family size and find that marginal rates should be lower (or, for the highest type in a nonlinear scheme, the same) for families

One issue that has received special attention is the treatment of second earners, typically women in two-parent families (often with children), who some suggest to have a higher elasticity of labor supply than primary earners or single individuals. It is argued that, accordingly, they should face lower marginal tax rates. See, for example, Boskin and Sheshinski (1983), Feldstein and Feenberg (1996), and Rosen (1976), and also Schroyen (2003) who considers intrahousehold behavior.

Second, incentives to marry are arguably affected by the tax system. In the United States, for example, married couples face a higher overall schedule than would be faced by two single individuals each earning half the couple's income, which creates a marriage penalty. On the other hand, if two individuals with highly unequal income marry, they face lower taxes on account of rate graduation, and this effect may exceed the first, creating marriage subsidies. Other provisions, especially those in transfer programs, can create relatively significant tax differentials as a consequence of the decision to marry.[140] See, for example, Alm and Whittington (1996) on the extent of marriage taxes and subsidies in the United States. Interestingly, the penalty effect is caused by a higher schedule often motivated by a desire to account for economies of scale in a manner that the analysis of subsection 10.3 indicates may not be optimal, and the subsidy effect is due to rate graduation that the analysis in subsection 3.4 indicates may not be optimal.

A common view is that the tax system should be marriage neutral to avoid distortion of the marriage decision. However, there may be positive externalities to marriage, especially when children are involved, in which case a subsidy may be optimal. It is often feared that eliminating marriage penalties or increasing subsidies, such as by making the schedules for married couples more generous, would be expensive and regressive (since the highest absolute penalties are borne by higher-income individuals), but just as in the case of adjustments for family size (see subsection 10.5), this view is misleading: It is possible to adjust the level of other schedules (here, raising that on single individuals) and the shape of all schedules (preserving the average amount paid at any level of income), and, regarding regressivity, one cannot define distributive goals meaningfully without first specifying the relative positions of various family units.

Third, the treatment of children in tax and transfer programs may influence procreation decisions, a possibility that is usually raised with respect to transfer programs.[141] Nearly all systems (except in countries that specially penalize or forbid additional births) make the net treatment at least somewhat more generous as the number of children increases, and the analysis in subsection 10.5 suggests the possibility that significant adjustments may be optimal. Whether any effect of such treatment on childbearing is desirable depends on the net external effects of having children. This problem

with more children. An explanation for this result can be found in the analysis of subsection 10.5, assuming that the absolute amount of adjustments required to equalize effective marginal utilities across family types is rising with income, as seems plausible.

[140] For a survey of evidence on the effects of welfare on family structure, see Moffitt (1992).

[141] Regarding the effect of welfare on fertility, see Schultz (1994). Whittington, Alm, and Peters (1990) offer evidence that income tax exemptions affect the birth rate.

is particularly vexing when one of the externalities (under at least some views) pertains to the births per se, i.e., the positive externality to the child thus created. See, for example, Mirrlees (1972b) and Nerlove, Razin, and Sadka (1986).

11. Tax administration and enforcement

Much analysis of taxation, as reflected in the foregoing sections, abstracts from concerns about administration and enforcement, implicitly assuming that tax liability is ascertained accurately and collected costlessly, without any resistance by taxpayers. This view, of course, is unrealistic. For example, in the United States, public and private collection costs for the income tax are approximately 10% of revenues, and it is estimated that over 15% of tax liability is unpaid.[142]

As outlined by Stiglitz (1985a), tax avoidance—generally understood as constituting reduction in tax obligations through manipulations permitted by law—comes in a number of basic forms: deferral (moving a nominal tax liability to a later date, thereby capturing the interest in the interim), tax arbitrage between individuals or entities with different marginal tax rates (including the case of the same taxpayer, facing different rates in different periods), and arbitrage across earnings flows facing different tax treatment (for example, dividends versus interest for firms, ordinary income versus capital gains for individuals). If unconstrained by capital market imperfections or legal limits (such as on the ability to deduct capital losses), Stiglitz suggests that individuals could eliminate their tax liability altogether without changing their underlying real behavior. Gordon and Slemrod (1988) calculated that the United States tax system in 1983 collected, roughly, no revenue from the taxation of capital income. As a result of subsequent events, Gordon, Kalambokidis, and Slemrod (2004) find that substantial revenue was raised from the taxation of capital in 1995, but Gordon et al.'s (2004) update to account for 2003 legislation concludes that, closer to the previous situation, little such revenue is collected as of 2004.

Tax evasion—illegal nonpayment—is also significant, as reflected in the aforementioned estimate of unpaid income tax liability. Although nearly 100% voluntary compliance is achieved on wages and salaries (which are subject to information reporting requirements), voluntary reporting is barely over 40% for self-employment income, due in large part to the difficulty of identifying evasion in the cash sector. It is suspected that evasion in many countries, especially developing economies, is much worse than in the United States, in significant part because so much more economic activity is in the latter category. It is worth noting that, on average, evasion tends to be worse as a percentage of income at the bottom of the income distribution whereas avoidance tends to be more significant at the upper end, in both cases on account of differential opportunities for tax reduction. (There is some important overlap; notably, the self-employed tend to avoid and evade more than average.)

[142] See, for example, Guyton et al. (2003), Internal Revenue Service (1996, 2005), and Slemrod (1996b).

The substantial literatures that measure avoidance and evasion, examine the effects of different policies on the extent of revenue loss, and analyze what policies are optimal in light of these problems are largely beyond the scope of this chapter but are surveyed elsewhere.[143] This section will focus on conceptual and normative issues that pertain most directly to the topics addressed in previous sections.

11.1. Choice of tax systems

Most analyses simply assume that some tax instruments are available and others are not (just commodity taxes, commodity taxes and income taxes, only linear income taxes, no ability taxes).[144] However, as emphasized by Mirrlees (1971), Atkinson and Stiglitz (1980), and Slemrod (1990), among others, it is important that the presumed set of available instruments be motivated by administrative and enforcement concerns that indicate what actually is feasible. Ideally, these concerns would not be stipulated but rather would be made endogenous, which is the point of much of the literature examined in subsection 11.2. Often, feasibility is a matter of degree, and one must choose among various imperfect systems, the quality of each being determined by policy choices regarding administration and enforcement and also by how the instrument is used (e.g., the extent of evasion may depend on rates and on what other taxes are in place).

It is useful to consider a few examples of how administrative considerations may affect the choice among tax systems. As noted in subsection 3.5.1, Stern (1982) compared an imperfect ability tax (combined with a linear income tax) to a perfect (that is, error-free) nonlinear income tax. More broadly, the analysis in that section considered, for any signals that may correlate with ability, how one might optimally design a nonlinear income tax system that could be made a function of the signal. Further analysis would make endogenous both the government's categorization (how accurately to observe the signal, how to set burdens of proof) and private efforts to manipulate such signals.

In comparing a linear and a nonlinear income tax (subsections 3.2 and 3.4), it is obvious that the latter dominates the former in a world without administration costs and avoidance activity. However, when all income is taxed at the same rate, substitute taxation—notably, collection at the source—becomes feasible and may have compliance advantages. Furthermore, incentives to engage in transactions to shift income between taxpayers subject to different marginal rates are eliminated. Hence, if the welfare gain from nonlinear taxation is modest, linear income taxation may be preferable.

As a further illustration, consider the choice between income taxation and broad commodity taxes, like sales taxes or a VAT. Setting aside cases in which differential

[143] See, for example, Andreoni, Erard and Feinstein (1998), Cowell (1990), Roth, Scholz, and Witte (1989), and Slemrod and Yitzhaki (2002).

[144] Similar assumptions are invoked concerning key elements of the tax base, such as whether nonpecuniary income (e.g., imputed rent from owner-occupied housing) may be taxed (see subsection 12.1.1) or whether capital gains can only be taxed on a realization basis (see subsection 9.4).

commodity taxation would be optimal, it would seem that commodity taxes are redundant and, if they have any additional administrative cost, undesirable. However, some propose that such taxes may be useful if there is significant income tax evasion in some sectors. Due to its method of collection, many believe that a VAT is overall harder to evade, so shifting significant collections away from the income tax may be advantageous. Moreover, even if commodity taxes are subject to significant evasion (perhaps in the same areas as income taxation, such as those involving informal activity conducted largely in cash), it has been further suggested that a combination of the two systems may be optimal. The intuition is that individuals who evade the income tax would still pay taxes on their consumption (in most sectors). Yet as Kesselman (1993) shows, when general equilibrium effects on prices and wages are taken into account, this idea is incorrect in the case in which commodity taxes are fully evaded in the same sectors as those in which the income tax is evaded—an approximately plausible scenario since commodity tax evasion typically will be both possible in such cases and also necessary to avoid detection of the income tax evasion. The reason is that the ultimate incidence is the same regardless of whether a tax is levied on a producer's inputs (in particular, labor) or sales.[145]

These few examples and the limited research to date suggest that greater attention to the choice among tax systems is warranted. Whether or not to have a 20% VAT, relying far less on income taxes, is probably a more important decision than how to set commodity tax differentials in subtle ways in light of the qualifications to the uniformity result noted in subsection 4.3. Such system choices are likely to be particularly important for developing countries, where fewer options are feasible and the available instruments are changing over time (and in ways that are influenced by other government policies).[146]

11.2. Optimal administration and enforcement

The determination of optimal administration and enforcement of a given tax system is itself complex. First, there are many dimensions, ranging from the design of tax rules to the intensity of audits, extent of information reporting requirements, accuracy of adjudication of disputes, setting of penalties, and allocation of resources across types of taxes and taxpayers. Second, problems of avoidance and evasion and the responses thereto have important feedback effects, notably, on what tax rates are optimal (should the optimal rate on a commodity subject to evasion be higher or lower?) and on the initial choice of which forms of taxation to employ.[147]

[145] Kesselman does find some benefit to shifting toward commodity taxation when evasion of the latter (unlike evasion of the income tax) is incomplete, but he argues that for plausible parameter values this benefit is small. Boadway, Marchand, and Pestieau (1994) find supplementation with commodity taxation to be desirable in a model in which it, unlike income taxation, is not subject to evasion.

[146] See Gordon and Li (2005).

[147] Although private compliance and avoidance activity is considered here, no attention will be given to the possibility that it may be optimal to regulate such activity directly, for example, by taxing professional tax advice because its use may be socially excessive. See Kaplow (1998).

The problem is also challenging because of the subtlety of the harm due to avoidance and evasion, which involves a transfer to taxpayers (or, one might say, a lack of a transfer to the government) rather than a destruction of resources.[148] To see this, suppose that there exists only a tax on a single commodity and that taxpayers are costlessly able to avoid taxation on half of their purchases. Here, it would not make sense to expend governmental resources to reduce avoidance, for instead the government could simply double the tax rate. Although this example is artificial, it illustrates that the costs of nonpayment are less obvious than they may first appear. To further complicate matters, consider the fact that, while raising rates tends to increase distortion, so does increased enforcement since, if it works, it raises effective rates, which are the source of distortion. In addition, increasing enforcement entails direct resource costs.

To appreciate the benefits of enforcement and to determine optimal policy, it is necessary to return explicitly to the SWF; indeed, the problem of administration was one of the motivations offered in subsection 2.3 for making direct use of the SWF. The central idea is that, whatever are the criteria used to determine whether one or another tax system or level of tax rates is optimal in the first instance, the same criteria should be employed to assess deviations and efforts to correct them.

This approach has been followed increasingly in recent work, including Kaplow (1990, 1998), Mayshar (1991), Slemrod (2001), and Slemrod and Yitzhaki (1987, 1996). The models employed vary in abstraction (whether they consider generally instruments available to the government or particular means of enforcement) and in the private behavior addressed. For example, Kaplow (1990) considers optimal government expenditure on enforcement that increases the fraction of taxpayers observed by the tax authority and how the optimum depends on private evasion reactions; Mayshar (1991) examines a tax authority that chooses a range of policy instruments where taxpayers choose the amount of labor effort to devote to sheltering; and Slemrod (2001) models private avoidance when increasing labor income increases avoidance opportunity. In this literature, social costs and benefits depend on how avoidance, evasion, and enforcement affect the equitable allocation of tax burdens across individuals, the extent of distortion caused by taxes, the amount of resources devoted to private compliance and tax reduction activity, government expenditures on administration, and risk-bearing.[149] It is useful to elaborate on some of these effects.

Tax equity is implicated because the achievable level of social welfare depends on whether individuals pay the correct amount of tax. After all, were this not a concern, society could rely on uniform lump-sum taxes and not have to worry about distortion. In reality, imperfections in defining the tax base and in administering the law combined

[148] Compare Shavell (1991) on theft.

[149] Earlier work on evasion and avoidance was largely positive. Allingham and Sandmo (1972) considered a setting like that in Becker's (1968) model of law enforcement; the taxpayer's only decision was the choice of how much to underreport income, and this choice depended on risk preferences. Optimal enforcement thus addressed risk-bearing concerns, as in Polinsky and Shavell (1979). Extensions to Allingham and Sandmo include Yitzhaki (1974, 1987). On models of law enforcement, see generally Polinsky and Shavell (2007).

with taxpayers' efforts to minimize tax obligations result in a system where mismeasurement often occurs. The welfare effect can be determined directly from the SWF. Standard first-order conditions—see expressions (3.8) for the linear income tax and (3.9) for the nonlinear income tax—have a $W'u_c$ term, indicating the marginal social value of a dollar to each taxpayer (the marginal utility to the taxpayer times the marginal social welfare weight). Holding constant the revenue to be raised from a particular group of taxpayers, greater error is associated with a greater dispersion in treatment.[150] Because u_c is strictly concave and, if the SWF is not utilitarian, W' is as well, mismeasurement, however produced, tends to be welfare-reducing. As developed in Kaplow (1998), the welfare cost is roughly given by a risk premium (determined, in the utilitarian case, by taxpayers' risk preferences, reflected by the curvature of their utility functions). One implication is that, for a given absolute error, mismeasurement is more costly to social welfare when it affects lower-income taxpayers on account of decreasing absolute risk aversion (and to a greater extent the more concave are their utility functions and the SWF).

A further point about potential inequity also bears on distortion. Specifically, as Bittker (1979) and Bradford (1980, 1986) argue, inequities often turn into inefficiencies. For example, if flight attendants are tax exempt on the value of free air travel or if sellers in the underground economy are effectively exempt from tax (see Kesselman (1989)), equilibrium wages and prices will adjust so that, for the marginal taxpayer, there is no gain, the benefit being passed on to consumers (or others). Accordingly, many imperfections in the tax system involve little inequity but do cause inefficiency. The same point holds regarding so-called tax shelters. For example, low-income housing tax shelters or tax-advantaged shopping center or office building developments are likely to result in greater investment and lower rents in the targeted areas of activity rather than windfalls to those who invest in the shelters.

As the foregoing point suggests, much avoidance and evasion can be analyzed by analogy to changes in marginal tax rates, such as for particular commodities. This approach is followed in Kaplow (1990, 1998). In some instances, one might be able to offset the distortion by changing explicit rates. But if the sector is underground or if evasion is selective, this will not be feasible. Instead, resources may need to be spent to reduce noncompliance, such as by expanding information-reporting, increasing auditing, and so forth. To the extent that such enforcement succeeds, effective tax rates will be driven higher. As noted previously, however, this may appear to increase distortion. There are two main reasons why such efforts may nevertheless be efficient. First, in some instances distortion may fall directly. For example, if most activity is taxed, exempting only some activity increases distortion through inefficient substitution. Second, one must make appropriate comparisons. Consider the prior hypothetical examples, in which the choice was between raising nominal rates and increasing enforcement. If, more realistically, some pay the full nominal rate and others pay less, then raising enforcement tends to be less distortionary. The reason relates to the familiar point that

[150] Compare Kaplow's (1989) analysis of horizontal inequity.

marginal distortion rises with the effective tax rate. Accordingly, subjecting some individuals to high effective tax rates and others to low (or zero) effective tax rates, as occurs when there is selective evasion, results in greater distortion than when everyone is subject to an intermediate tax rate, with which there is perfect compliance.

The analysis of the potential equity and efficiency effects of evasion indicates how the social costs can, in principle, properly be measured. Then it is possible to assess whether increased expenditures on one or another enforcement instrument are optimal. However, the analysis is further complicated by the fact that, in addition to the government's expenditures, private costs must also be considered. Note that, in general, it is possible for these costs—which unlike nonpayment of taxes are real resource costs—to fall or rise as enforcement increases. On one hand, some individuals may be induced to reduce avoidance and evasion activities because they are rendered unprofitable. On the other hand, some individuals may spend more, in order to continue to keep their income out of the government's hands.

Finally, it is useful to revisit briefly the question of how these issues bear on the setting of tax rates and the choice of tax systems. Regarding the former, the results are ambiguous. Higher nominal rates—such as in the extreme, initial example when evasion could be costlessly offset—may be optimal. However, if some individuals face the full nominal rate and others face, say, a zero rate, and little can be done to combat this, then lower rates may be optimal than otherwise (or a zero rate, when there are fixed costs and little tax can be collected), with greater reliance on taxes that have a more uniform effect on different taxpayers. Regarding tax systems, obviously those that are very costly to administer and highly imperfect even after enforcement is optimized are less attractive. Of course, all tax systems suffer in varying degrees, so the question is a comparative one.

11.3. Elasticity of taxable income

Important recent work motivated by problems of administration and enforcement examines what has come to be referred to as the elasticity of taxable income. The first-order conditions for the optimal tax problem—such as expressions (3.8) and (3.9)—depend on the elasticity of labor supply. It has been emphasized, however, that taxpayers respond to income taxes in many ways: reducing labor supply, shifting compensation to tax-preferred fringe benefits, making use of tax shelters, and evading outright. Furthermore, these responses all have a qualitatively similar effect on distortion. See, for example, Feldstein (1999) and Slemrod (1998).[151] One implication of this literature has been an upward revision in assessments of the distortionary cost of labor income taxation because estimates of the elasticity of taxable income are greater than estimates of

[151] There are qualifications, such as when taxpayers increase charitable contributions or invest in low-income housing tax shelters, activities that may be preferred because of the positive externalities that they produce.

the elasticity of labor supply.[152] Interestingly, one reason labor supply responses to tax reforms may be low is that taxpayers are able to respond on these other margins.

Another implication of the literature is to reinforce the importance of examining the choice of tax systems, tax rates, and tax enforcement parameters as part of a unified optimization. As Slemrod and Kopczuk (2002) argue, the pertinent elasticity, of taxable income, unlike the elasticity of labor supply, is in significant part determined by policy rather than a manifestation of individuals' exogenous preferences. For example, greater problems of avoidance and evasion—implying a higher elasticity of taxable income—may favor lower tax rates. But it is also true that higher optimal tax rates warrant greater expenditures to reduce avoidance and evasion. Slemrod and Kopczuk examine a model in which administrative cost considerations are the impediment to a comprehensive income tax base, and they find that a social desire for a more redistributive tax should be accompanied by greater administrative expenditures to broaden the base; conversely, the more costly it is to expand the base, the less redistribution is optimal. Kopczuk (2005) analyzes the base broadening of the 1986 Tax Reform Act in this light.[153]

12. Additional features of tax systems

Numerous particular features of tax systems have been studied in varying degrees of depth—far too many to mention, much less summarize, in a single, conceptual survey on taxation. Some elements, however, are particularly significant and have a close relationship to themes pursued in previous sections (especially section 11 on administration and enforcement), so they will be examined briefly here.

12.1. Tax base

12.1.1. Exclusion of nonpecuniary income

Idealized income and consumption tax systems envision a comprehensive base because it is thought to provide a better measure for purposes of distributive equity and because omissions generally are distortionary. There are, however, certain systematic exclusions that are often justified on account of the infeasibility—or significant administrative difficulty—of measurement. One of the most important sets of exclusions involves nonpecuniary sources of income.

[152] See, for example, Auten and Carroll (1995, 1999), Feldstein (1995), Giertz (2004), Goolsbee (2000), Gruber and Saez (2002), Moffitt and Wilhelm (2000), and Slemrod (1996a).

[153] See also Wilson (1989) and Yitzhaki (1979) who consider the optimal base of a commodity tax where base broadening reduces distortion but adding more commodities to the base is costly, and Weisbach (2000) who considers, by analogy to the commodity tax base problem, where to draw lines between taxable and exempt activities or transactions, recognizing that expanding the base may or may not be efficient depending on whether the newly included activity is a closer substitute for taxed or tax-exempt activities.

The most fundamental such exclusion is the value of leisure, sometimes referred to as imputed income due to household services.[154] Indeed, the central distortion caused by income taxation (and consumption taxation, where the pertinent exclusion pertains to the value of nonmarket time)—the labor-leisure distortion—is directly attributable to this exclusion. See subsection 3.5.1 on ability taxation.

Another important exclusion, related to the topic of the next subsection, involves nonpecuniary features of market employment. These include both positive features—air-conditioning or artwork—and negative attributes—unpleasantness and danger associated, for example, with mining, harvesting, and work on assembly lines. Standard labor market theory suggests that nonpecuniary features of employment will, in equilibrium, be offset by compensating wage differentials. See Rosen (1986). However, the compensating differentials are subject to taxation—higher wages are taxed and wage reductions are implicitly excluded because never earned—whereas the offsetting amenities for which they compensate are not recognized by the tax system. Hence, omitting nonpecuniary job characteristics when measuring labor income or consumption distorts workplace attributes and the allocation of labor across jobs.

An additional significant exclusion, one particularly relevant to an income tax (that reaches capital as well as labor income), is of the imputed rent from consumer durables, most importantly, housing. Rental services, a form of consumption, are not deductible. But if one owns durables rather than renting them, there is no tax on the imputed rent. Put another way, the return to capital is, in principle, subject to tax, but if the return is in the form of services to oneself, for which no explicit rent is paid, the return is effectively exempt.[155] Because housing alone is such a large fraction of the capital stock, this exclusion is hardly innocuous.

12.1.2. Business versus personal expenditures

Related to the preceding subsection's discussion of nonpecuniary features of employment, there is a more general problem of distinguishing business (or, more broadly, income-producing) and personal expenditures. Pure costs of doing business (a sole proprietor's cost of goods sold, rent, utility bills, and so forth) must be deducted or otherwise excluded in properly measuring net income, whereas items of consumption, which may be heavily present in many fringe benefits, need to be kept within the tax base of an income or consumption tax.

[154] This latter characterization is potentially misleading because one might mistakenly assume that only labor used in direct household production (meal preparation, cleaning, home repair, child care) is relevant whereas mismeasurement and distortion are involved on account of the omission of any non-market time from the tax base.

[155] Observe that this exclusion is not directly attributable to the home mortgage interest deduction. Under an ideal income tax, all interest payments are in principle deductible. (They are negative interest receipts, which are taxable.) The benefit of untaxed imputed rent is fully available to an owner-occupier who has no debt. The primary relevance of the mortgage interest deduction is to make the benefit available to individuals with insufficient net worth to own outright.

Some measurement problems concern pure subterfuge, such as when an individual attempts to deduct a large fraction of housing costs as a "home office" although little work is performed in the home or when employers pay for what are tantamount to employee vacations. Others involve myriad situations of genuinely mixed use: Free meals for restaurant employees help monitor quality and make servers more informative to customers, entertainment may improve camaraderie or client relationships, and improved working conditions may simultaneously raise productivity and utility. One problem is distinguishing the former and latter cases. A second is determining optimal treatment in cases in which consumption and production are intermingled, on which see Katz and Mankiw (1985). Observe that, in either situation, to the extent that there are utility benefits to workers, we would expect wages to adjust, so ultimately the problem is one of distortion, here involving forms of expenditure at the workplace and the choice of occupations.

12.1.3. Retirement savings[156]

Under an accrual income tax (in contrast to a labor income tax or a consumption tax), the return to savings is included in the tax base. See subsection 5.1.1. A common feature of income tax systems, such as in the United States, is to provide tax preferences for retirement savings through employers (pension plans) and individual retirement accounts of various sorts. Such schemes typically provide consumption tax treatment by allowing an exclusion or deduction from current income for contributions, permitting tax-free build-up, and subjecting withdrawals to tax.

Retirement savings provisions are variously rationalized on the ground that they move in the direction of a consumption tax, deemed to be preferable; that they increase national savings (which depends on the empirical question of whether the income or substitution effect dominates), believed by some to be desirable; or that they offset individuals' tendency to provide inadequately for their retirement (see subsection 5.3 on social security). The actual implementation of such schemes—particularly employer pension plans, which are heavily regulated—is complex. Furthermore, it is uncertain the extent to which the latter objectives are achieved. Notably, much of what is contributed to retirement savings plans may not be additional savings but shifts of funds that would have been saved in any event, and, regarding paternalism, myopic individuals may well be those least likely to respond to savings incentives.

12.1.4. Tax expenditures

Departures from a tax base that involve exclusions, deductions, or other preferences are sometimes generically referred to as tax expenditures, a view championed by Surrey (1973). The notion is that granting special tax treatment—e.g., for expenditures on

[156] See generally Bernheim (2002) on taxation and saving.

energy conservation—is tantamount to a direct budget outlay for the activity. Surrey further proposed that an annual tax expenditure budget be compiled (as it now is), to provide accountability for such expenditures that, as a whole, constitute a sizeable fraction of government spending. Additionally, he generally opposed tax expenditures' existence on grounds of accountability and their "upside-down" effect, being worth more to taxpayers in higher brackets.

Although there clearly is some virtue to this viewpoint, there are some difficulties as well.[157] One concerns what constitutes a tax expenditure, the argument often going to the merits of the choice of tax base. For example, favorable treatment of retirement savings is a tax expenditure under an income tax, but not under a consumption tax, where subjecting ordinary savings to tax is seen as tax penalized. Concerns about regressivity can, in principle, be met by tax rate adjustments, as exemplified in the 1986 Tax Reform Act in the United States, where the repeal of many tax expenditures was accompanied by a purportedly distribution-neutral adjustment of tax rates. Whether it is efficient to deliver subsidies through the tax system or otherwise and other arguments going to the desirability of various tax expenditures depends on analysis of pertinent specifics. Nevertheless, the basic point that there is no clear distinction between spending and selective tax reduction is important both for tax policy and a broader range of fiscal matters.

12.2. Forms of consumption taxation

The discussion of commodity taxation in section 4 and the comparison of income and consumption taxation in subsection 5.1.1 provide some insight into the nature of consumption taxation. It is useful, however, to explore variations in the form of consumption taxation, some of which are equivalent to others in principle but may differ with regard to administrability and evasion.

12.2.1. Cash-flow consumption taxation

If a uniform tax on all forms of consumption is desired—and possibly at different marginal rates depending on individuals' aggregate consumption—it is not necessary to measure each individual's expenditures on each and every commodity. Instead, one may employ cash-flow taxation, as developed by Andrews (1974). Because total consumption in an accounting period equals income minus net savings (i.e., minus deposits and plus withdrawals), a consumption tax base may be defined just as an income tax base (that includes labor and capital income), making an adjustment for net savings.

Indeed, implementing such a consumption tax is likely to be significantly easier than defining the income tax base, even though the former on its face requires an additional

[157] Much of the leading commentary on the tax expenditure concept is by legal academics. See, for example, Bittker (1969), Griffith (1989), Shaviro (2004), Surrey and McDaniel (1985), and Weisbach and Nussim (2004).

set of adjustments. The central reason is that many of the most difficult measurement problems in defining income involve capital income—especially accruals, such as with capital gains and depreciation (see subsection 9.4). But the consumption tax's adjustment for net savings makes many of these problems moot. For example, when an individual purchases an asset, a deduction would be allowed. Any intervening changes in value can be ignored under a consumption tax; all that is necessary is to include the proceeds upon ultimate disposition. Even that final step is unnecessary if there is reinvestment. Thus, for assets held in an account (say, a brokerage account or a closely-held firm), one need only track flows in and out; all changes in value within the account are irrelevant. For this reason, Andrews (1974), Bradford (1986), and others find a consumption tax superior to an income tax on administrative grounds. For further aspects of the comparison, see subsection 5.1.2.

As noted in subsections 4.2 and 5.1.1, a uniform consumption tax is equivalent to a labor income tax (in a basic setting with linear taxes), so another way to implement a consumption tax is to tax all labor income while exempting capital income. (Following the analysis of subsection 9.2.2, these systems are also equivalent in a world with uncertainty, taking into account individuals' portfolio adjustments.) In some respects, a labor income tax seems particularly easy to administer, for only wages need be taxed. However, disentangling wage income from capital income is sometimes difficult, notably for the self-employed, whereas consumption taxation, which only needs to track the outflow of funds to the owner, may be easier to implement. Hybrid schemes have also been proposed, both regarding sources of labor income and the treatment of other issues, such as purchases of consumer durables. See Bradford and U.S. Department of Treasury (1984).

12.2.2. VAT and sales taxation

The other main forms of consumption taxation—which generally must be linear, unlike a cash-flow consumption tax—are value-added taxes (VATs) and sales taxes. These can be imposed uniformly, or subject to rate variations and exemptions (such as preferences for expenditures on food), thereby implementing the full range of commodity tax schemes analyzed in section 4.

A VAT is applied to the value added at each stage of production. By contrast, a sales tax applies only to final sales to ultimate consumers. In either case, the same amount, in principle, is subject to tax, and the equilibrium incidence is the same.[158] The main differences are administrative.

[158] See generally Bradford (1996a, 1996b). Employing a sales tax can be administratively problematic when some goods are intermediate for some purchases and final for others. Additionally, some sales tax schemes are not legally defined in terms of the underlying principle of ultimate sales to consumers but instead purport to cover a broad range of sales subject to numerous specific exemptions covering most intermediate uses; however, this latter approach sometimes results in multiple taxation (referred to as "cascading") when there are gaps in the exemptions.

A sales tax requires policing only the final stage, but this stage is often difficult to monitor where there are large numbers of small retailers. A VAT covers all stages, but production and wholesale distribution are often more concentrated, making enforcement easier for much of the tax base. Additionally, certain forms of the VAT have a self-enforcing feature: Under the credit-invoice method, a seller at any stage pays tax on gross receipts and, in order to receive an offset for taxes previously paid by others on purchased inputs, the seller must produce invoices that confirm payment of tax at the prior stages.

13. Tax equity

Sections 3 through 12 consider how to employ various tax instruments to maximize a standard SWF. This section considers issues bearing on whether social welfare, conventionally understood, should be the sole social objective and on the form of the SWF.

13.1. Welfarism

Welfarism is the principle that social decisions should be based exclusively on how they affect individuals' utilities (welfare or well-being). Put another way, data on the levels of utility achieved by each individual under a policy are deemed to be sufficient information to ground social choice. It is common in optimal income tax analysis to employ a function of the additive form, displayed in expression (2.1), but that is not essential to the concept.

The motivation for the welfarist approach is not only that each individual's well-being should matter, but that anything independent of anyone's well-being should not. The appeal of welfarism is bolstered by the fact that any nonwelfarist approach conflicts with the Pareto principle; that is, following any nonwelfarist approach will sometimes favor a regime under which everyone would be worse off. See Kaplow and Shavell (2001).

An important clarification is that much of relevance under welfarism may be important because of its indirect, ultimate effects on individuals' utilities, but this does not make something a social value in its own right. Some considerations may serve as proxies for well-being; for example, a simplified tax system may be desirable because it is less costly to administer, which in turn saves resources, allowing a higher level of welfare to be achieved. Furthermore, certain factors may be components of well-being for some individuals; in this spirit, subsection 3.5.2 noted the possibility that individuals may have preferences regarding redistribution itself. Welfarism entails the view that once all such effects on well-being are taken into account, the relevance of any given consideration is exhausted.

Although welfarism has not been highly controversial among economists, the discussion in subsection 13.3 will indicate that a number of familiar normative criteria that have been used to assess tax policy do appear to be nonwelfarist, unless they are understood purely as proxies rather than as independent principles.[159] In more direct apparent confrontation, Sen (1985, 1997) has suggested that individuals' situations should be assessed based on their capabilities and functionings—a sort of list of means of fulfillment—rather than solely their well-being. Rawls's (1971, 1982) notion of primary goods has been seen by many in a similar light. A number of issues have been raised concerning these nonwelfarist approaches. First, what is on the privileged list and how is this determined, if capabilities, functionings, and primary goods are to be viewed not purely instrumentally but rather as constituents of the good itself? Second and related is the problem of relative weightings, which are necessary if there is more than one item on the list. See, for example, Blair (1988) and Gibbard (1979). Indeed, it is straightforward to demonstrate that these alternative approaches conflict with the Pareto principle, unless by chance the lists and weightings correspond precisely to those implicit in all individuals' utility functions (in which case there is no disagreement with the dictates of welfarism).[160] Underlying these problems is the question of why society should deviate from welfarism. Perhaps the appeal of these theories lies not in a genuine rejection of welfarism but, to the contrary, in a concern that well-being is often assessed too narrowly. Indeed, the appendices to Sen (1985) and some of his other work on development emphasize that frequently used measures like per capita GDP do not adequately capture well-being, whereas supplementation with additional factors provides a more accurate indicator.

As a practical matter, this debate about welfarism has had little impact on the economic analysis of tax policy. Independent of the merits of the dispute, most tax instruments are based on income, consumption, and related observable flows. Different views on the proper social objective matter primarily when individual differences are observable, such as when some individuals are disabled. In this realm, actual policy seems to reflect a mixed position. On one hand, accommodation requirements may be prompted by a desire to equalize capability rather than utility. On the other hand, this approach may be motivated by concerns for welfare, and there seems to be little enthusiasm for policies that consciously seek to enhance capabilities at the expense of beneficiaries' well-being—although this may occur implicitly if non-cost-justified accommodations are required in lieu of alternative forms of assistance that recipients would, all things considered, value more highly.

[159] Welfarism is highly controversial outside of economics. Notably, much of twentieth-century moral philosophy is critical of the approach. See, for example, Sen and Williams (1982), for competing views. For a survey, analysis, and response to nonwelfarist writings, see Kaplow and Shavell (2002).

[160] Suppose, for example, that all individuals have the same utility function, there are two goods, production is centralized, the marginal rate of transformation between the goods is 1, and the nonwelfarist theory deems the two goods to be of equal importance. The planner (taken here to be egalitarian) would produce equal amounts of each good and distribute them pro rata. If, however, individuals' utility functions are optimized at any other combination of the two goods, everyone would be worse off than if the planner selected utility-maximizing proportions of the two goods instead.

13.2. Choice of social welfare function

As noted in subsection 2.3, when an explicit SWF is required, notably, in optimal income tax simulations, an additive, often iso-elastic form is used. See expression (2.2). As discussed, the inequality parameter, e, may range from 0 (corresponding to a utilitarian summation) to ∞, the limiting case that corresponds to maximization of the utility of the least-well-off individual, inspired by Rawls (1971).

Although most work is formally agnostic about the concavity of the SWF, important arguments have been offered. Harsanyi (1953), predating Rawls's use of a "veil of ignorance" or "original position," postulated that the n individuals in a society had to choose regimes not knowing their actual identity, with each believing that there was a $1/n$ chance that he or she would be any of the n individuals in society. Harsanyi showed how each individual's expected utility in this setting corresponded to the utilitarian maximand. See also Vickrey (1945). Independently, Harsanyi (1955) developed an argument that assumed that each individual's utility followed the rationality axioms of decision (and expected utility) theory, likewise for the SWF, and that the SWF depended solely on individuals' utilities in a positive and symmetric fashion. From this, he deduced that the SWF had to be utilitarian.[161]

Additional, related arguments for a utilitarian SWF have been offered. Hammond (1983) demonstrated that no other SWF was time consistent. One way to express the idea is to note that, from an initial point, a reform involving uncertainty may be favored because it raises expected social welfare; however, after enactment, with a nonlinear SWF it is possible that repeal would be deemed optimal. Kaplow (1995) showed that, for any strictly concave SWF, one can construct examples in which a reform would be unanimously preferred ex ante by all individuals but rejected under the SWF; that is, there is a conflict with the Pareto principle.

Sen (1997) and others view utilitarianism as insufficiently egalitarian. The meaning of such an objection is not entirely clear. A utilitarian SWF is formally egalitarian (everyone counts equally, that is symmetrically or anonymously), and in simple cases without incentive concerns it favors complete equalization. In realistic settings, none of the standard SWFs favor complete equality; then, more concave SWFs favor greater equality. However, the degree of equality favored by any given SWF is a matter of subtlety and controversy, for it depends on individuals' utility functions (both the degree of concavity, which itself affects the optimal extent of redistribution, and the labor supply elasticity), the distribution of abilities, and in more complex models on many additional factors. Hence, it seems difficult to have an a priori view on the extent of inequality that should be tolerated, from which one might deduce the appropriate concavity of the SWF.

Rawls's (1971) maximin claim, translated into the present framework, is one of the few other specific SWFs that has been advocated. However, it has not commanded wide

[161] For a discussion of objections and responses, see, for example, Broome (1984), Diamond (1967), Harsanyi (1975), Myerson (1981), and Strotz (1958).

acceptance because of its extreme implication—that social welfare is raised by making nearly everyone in society miserable as long as there exists one slightly more miserable person who gains infinitesimally—and because it purports to be grounded in the original position and rationality assumptions that, per Harsanyi, imply utilitarianism. See, for example, Arrow (1973) and Hare (1973).

An additional set of issues surrounding the choice of SWF concerns whose welfare is to be included. Key dimensions include geographic scope (national versus international), the weighting of future generations, whether average or total welfare should be maximized (which is highly pertinent to issues bearing on population size), and whether the welfare of other sentient beings should count.

13.3. Other normative criteria

13.3.1. Traditional principles

Prior to the advent of modern welfare economics and its embodiment in optimal tax theory—and continuing to a lesser extent to the present—tax equity was judged by a range of criteria, including vertical and horizontal equity, ability to pay, the benefit principle, and principles of equal sacrifice. See Musgrave (1959, 1985). Additionally, certain definitions were sometimes treated as if they were normative criteria; notably, the Haig-Simons income definition was used as the foundation for articulating a comprehensive tax base that was portrayed as a normative ideal.[162]

Many of these principles can be understood as intuitive notions of distributive justice, or certain aspects thereof. Nevertheless, some appear to be limited to the funding of public goods—most obviously the benefit principle but also principles of equal sacrifice (with redistribution, it cannot be that everyone is sacrificing equally) and, under some interpretations, ability to pay. As noted in subsection 7.3 on benefit taxation, it is not apparent whether such notions have bite to the extent that redistributive taxation is also permitted. And if it is not, there is an arbitrariness (except under the benefit principle) due to the fact that the extent of permissible redistribution depends on the extent of public goods provision and on the distributive incidence of the public goods provided, which is a happenstance of technology and preferences. Likewise, the underlying basis for most of these principles is unclear.[163] (Exceptions are the benefit principle, which might be defended on libertarian, anti-redistributive grounds, and the equal marginal sacrifice version of the sacrifice principle, advanced by Edgeworth (1897) and Pigou (1928) as a corollary of utilitarianism.)

[162] Debates about the comprehensive tax base ideal and about the tax expenditure concept, see subsection 12.1.4, overlap.

[163] Blum and Kalven (1953) in a well-known essay employ a traditional approach and conclude that taxes should be proportional. As an indication of the difficulty of reasoning to particular conclusions in this fashion, Bankman and Griffith (1987) and Groves (1974), among others, have explained how Blum and Kalven rely largely on a presumption in favor of proportionality combined with broad skepticism toward many arguments concerning redistribution.

13.3.2. Horizontal equity

As stated by Musgrave (1959), horizontal equity is the principle that equals should be treated (specifically, taxed) equally. A number of approaches have been advanced for making this seemingly uncontroversial concept operational. See, for example, Aronson and Lambert (1994), Auerbach and Hassett (2002), Atkinson (1980), Feldstein (1976), King (1983), Musgrave (1990), and Plotnick (1981).

Two sets of difficulties have been identified. The first—recognized in much of the aforementioned literature developing indexes of horizontal inequity—concerns definitional problems. What if no two individuals are precisely equal? Relatedly, does it make a qualitative difference if individuals begin exactly equal or slightly unequal? Once measures are extended to unequals, as they have been, are they still measures of horizontal equity?

Second and more fundamental, just why is horizontal equity valued and why should society be willing to sacrifice social welfare, conventionally measured, in pursuit of horizontal equity? The measurement literature has said little on this question, which seems logically prior to deriving indexes since it is difficult to assess measurement instruments when the purpose of measurement is unclear. See Kaplow (1989). Kaplow (1995) shows that if any weight is given to horizontal equity, policies that make everyone better off may be rejected. This result is a special case of the more general subsequent demonstration of Kaplow and Shavell (2001), previously noted, that all nonwelfarist principles conflict with the Pareto principle. A plausible explanation for the strong concern about horizontal equity is that, although not itself a constituent of social welfare, violations serve as a proxy for factors associated with welfare reductions, such as greater inequality, risk-bearing, mistaken regulations, and abuse of power. See Kaplow (1989). After all, if individuals are truly equal in relevant respects, welfare maximization usually requires that they be treated equally.[164]

13.3.3. Inequality, poverty, progressivity, and redistribution

Related to the redistributive function of taxation, there have been developed various indexes of the extent of inequality and poverty existing in a society (whether before or after taking into account the effects of taxation) and of the degree of progressivity and redistribution attributed to all or part of the fiscal system.[165] These measures are sometimes employed to offer a normative assessment of taxation, the standard implication being that systems resulting in less inequality and poverty and, correspondingly, involving more progressivity and redistribution are superior.[166]

[164] Not always because, for example, there may be nonconvexities, as Stiglitz (1982b) demonstrates.

[165] On inequality, see Atkinson and Bourguignon (2000), Cowell (1995), Lambert (2001), and Silber (1999). On poverty, see Atkinson (1987b), Clark, Hemming, and Ulph (1981), Lambert (2001), Ravallion (1994), Ruggles (1990), Sen (1976), and Silber (1999). On progressivity and redistribution, see Jakobsson (1976), Kakwani (1977), Lambert (2001), Musgrave and Thin (1948), and Suits (1977).

[166] The indexes also have descriptive uses, which raise different issues. See Kaplow (2005).

There are two related difficulties with this approach. See Kaplow (2005). First, the implicit assumption that more is better is incorrect, for—as with much in economic policy—it is the optimal extent of redistribution rather than the maximal extent that is desired. Second, for a normative measure to be well-grounded, it must be derived from an SWF, as originally suggested by Dalton (1920) and undertaken by Atkinson (1970) with regard to the measurement of inequality. However, in thus deriving a measure, it is necessary as a prerequisite both to choose an SWF and to employ it to measure the level of social welfare under the regime in question. Because it is possible to derive the indexes in question only after a complete welfare assessment has already been obtained, it is difficult to see how the measure—of inequality, poverty, progressivity, or redistribution—can be of further normative use. In sum, indexes of inequality and poverty seem aimed at a component of social welfare, and measures of progressivity and redistribution at traits of policies that affect social welfare; hence, all are best seen, as many of the other normative criteria surveyed here, as proxies for welfare rather than as ultimate bases for social evaluation.

14. Conclusion

This essay has offered a conceptual survey of taxation. It illustrates how many forms of taxation and widely varied issues of tax policy are illuminated by relating the analysis to a central, unifying framework. Specifically, the model of optimal income taxation, extended to incorporate commodity taxes, serves as the foundation for understanding most of the subjects considered, including government expenditures on transfers and public goods and the use of taxes and other instruments to control externalities. Furthermore, grounding normative assessment explicitly in the welfare economic framework—where necessary making reference to a social welfare function or, by holding distribution constant, judging comparisons through use of the Pareto principle—renders policy evaluation more consistent and cogent.

Future research could advance the mission of providing a more integrated view of various elements of taxation. Specifically, the existing body of tax research, far too vast to examine here, could be better appreciated and its development more precisely guided if its relationship to core principles and structures was more often made explicit. In addition, the central building blocks, including the income tax itself, would benefit from further study because, despite the difficulty of optimal income tax analysis and what may appear to be the near-exhaustion of basic extensions, even slight advances have potentially great payoffs. Empirical research on taxation, which is beyond the scope of this survey, also could profit from the sharper definition of pertinent issues that would flow from the foregoing research program.

Acknowledgements

I am grateful to Alan Auerbach, James Hines, and Steven Shavell for comments and the John M. Olin Center for Law, Economics, and Business at Harvard University for finan-

cial support. Further elaboration on a number of the subjects addressed herein appears in *The Theory of Taxation and Public Economics* (Kaplow, in press).

References

Aaron, H.J. (1975). Who Pays the Property Tax? Brookings Institution, Washington, D.C.

Aaron, H.J. (Ed.) (1976). Inflation and the Income Tax. Brookings Institution, Washington, D.C.

Aaron, H.J., McGuire, M. (1970). "Public Goods and Income Distribution". Econometrica 38, 907–920.

Aaron, H.J., McGuire, M. (1976). "Reply to Geoffrey Brennan, 'The Distributional Implications of Public Goods' ". Econometrica 44, 401–404.

Aaron, H.J., Munnell, A.H. (1992). "Reassessing the Role for Wealth Transfer Taxes". National Tax Journal 45, 119–143.

Acs, G., Coe, N., Watson, K., Lerman, R.I. (1998). "Does Work Pay? An Analysis of the Work Incentives under TANF", Occasional Paper No. 9. The Urban Institute, Washington, D.C.

Akerlof, G.A. (1978). "The Economics of 'Tagging' as Applied to the Optimal Income Tax, Welfare Programs, and Manpower Planning". American Economic Review 68, 8–19.

Allen, F. (1982). "Optimal Linear Income Taxation with General Equilibrium Effects on Wages". Journal of Public Economics 17, 135–143.

Allgood, S., Snow, A. (1998). "The Marginal Cost of Raising Tax Revenue and Redistributing Income". Journal of Political Economy 106, 1246–1273.

Allingham, M.G., Sandmo, A. (1972). "Income Tax Evasion: A Theoretical Analysis". Journal of Public Economics 1, 323–338.

Alm, J., Wallace, S. (2000). "Are the Rich Different?". In: Slemrod, J.B. (Ed.), Does Atlas Shrug?: The Economic Consequences of Taxing the Rich. Russell Sage Foundation, New York, pp. 165–187.

Alm, J., Whittington, L.A. (1996). "The Rise and Fall and Rise ... of the Marriage Tax". National Tax Journal 49, 571–589.

American Law Institute (1989). Federal Income Tax Project, Subchapter C (Supplemental Study, Reporter's Study Draft by W.D. Andrews). American Law Institute, Philadelphia.

American Law Institute (1993). Federal Income Tax Project, Integration of the Individual and Corporate Income Taxes (Reporter's Study of Corporate Tax Integration by A.C. Warren). American Law Institute, Philadelphia.

Andreoni, J. (1990). "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving". Economic Journal 100, 464–477.

Andreoni, J. (2006). "Philanthropy". In: Kolm, S.-C., Mercier Ythier, J. (Eds.), Handbook of the Economics of Giving, Altruism, and Reciprocity, vol. 2. North-Holland, Amsterdam, pp. 1201–1269.

Andreoni, J., Erard, B., Feinstein, J. (1998). "Tax Compliance". Journal of Economic Literature 36, 818–860.

Andrews, W.D. (1974). "A Consumption-Type or Cash Flow Personal Income Tax". Harvard Law Review 87, 1113–1188.

Andrews, W.D., Bradford, D.F. (1988). "Savings Incentives in a Hybrid Income Tax". In: Aaron, H.J., Galper, H., Pechman, J.A. (Eds.), Uneasy Compromise: Problems of a Hybrid Income-Consumption Tax. Brookings Institution, Washington, D.C., pp. 269–300.

Apps, P.F., Rees, R. (1988). "Taxation and the Household". Journal of Public Economics 35, 355–369.

Aronson, J.R., Lambert, P.J. (1994). "Decomposing the Gini Coefficient to Reveal the Vertical, Horizontal, and Reranking Effects of Income Taxation". National Tax Journal 47, 273–294.

Arrow, K.J. (1973). "Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice". Journal of Philosophy 70, 245–263.

Atkinson, A.B. (1970). "On the Measurement of Inequality". Journal of Economic Theory 2, 244–263.

Atkinson, A.B. (1980). "Horizontal Equity and the Distribution of the Tax Burden". In: Aaron, H.J., Boskin, M.J. (Eds.), The Economics of Taxation. Brookings Institution, Washington, D.C., pp. 3–18.

Atkinson, A.B. (1987a). “Income Maintenance and Social Insurance”. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 2. North-Holland, Amsterdam, pp. 779–908.

Atkinson, A.B. (1987b). “On the Measurement of Poverty”. Econometrica 55, 749–764.

Atkinson, A.B., Bourguignon, F. (Eds.) (2000). Handbook of Income Distribution, vol. 1. Elsevier, Amsterdam.

Atkinson, A.B., Sandmo, A. (1980). “Welfare Implications of the Taxation of Savings”. Economic Journal 90, 529–549.

Atkinson, A.B., Stern, N.H. (1974). “Pigou, Taxation and Public Goods”. Review of Economic Studies 41, 119–128.

Atkinson, A.B., Stiglitz, J.E. (1972). “The Structure of Indirect Taxation and Economic Efficiency”. Journal of Public Economics 1, 97–119.

Atkinson, A.B., Stiglitz, J.E. (1976). “The Design of Tax Structure: Direct versus Indirect Taxation”. Journal of Public Economics 6, 55–75.

Atkinson, A.B., Stiglitz, J.E. (1980). Lectures on Public Economics. McGraw-Hill Book Company, New York.

Auerbach, A.J. (1979). “Share Valuation and Corporate Equity Policy”. Journal of Public Economics 11, 291–305.

Auerbach, A.J. (1985). “The Theory of Excess Burden and Optimal Taxation”. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 1. North-Holland, Amsterdam, pp. 61–127.

Auerbach, A.J. (1991). “Retrospective Capital Gains Taxation”. American Economic Review 81, 167–178.

Auerbach, A.J. (2002). “Taxation and Corporate Financial Policy”. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 3. Elsevier, Amsterdam, pp. 1251–1292.

Auerbach, A.J., Bradford, D.F. (2004). “Generalized Cash-Flow Taxation”. Journal of Public Economics 88, 957–980.

Auerbach, A.J., Feldstein, M. (Eds.) (1985). Handbook of Public Economics, vol. 1. North-Holland, Amsterdam.

Auerbach, A.J., Feldstein, M. (Eds.) (1987). Handbook of Public Economics, vol. 2. North-Holland, Amsterdam.

Auerbach, A.J., Feldstein, M. (Eds.) (2002a). Handbook of Public Economics, vol. 3. Elsevier, Amsterdam.

Auerbach, A.J., Feldstein, M. (Eds.) (2002b). Handbook of Public Economics, vol. 4. Elsevier, Amsterdam.

Auerbach, A.J., Hassett, K.A. (2002). “A New Measure of Horizontal Equity”. American Economic Review 92, 1116–1125.

Auerbach, A.J., Hines, J.R. Jr. (2002). “Taxation and Economic Efficiency”. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 3. Elsevier, Amsterdam, pp. 1347–1421.

Auerbach, A.J., Kotlikoff, L.J. (1987). Dynamic Fiscal Policy. Cambridge University Press, Cambridge.

Auerbach, A.J., Gokhale, J., Kotlikoff, L.J. (1991). “Generational Accounts: A Meaningful Alternative to Deficit Accounting”. In: Bradford, D. (Ed.), Tax Policy and the Economy, vol. 5. MIT Press, Cambridge, Mass., pp. 55–110.

Auten, G., Carroll, R. (1995). “Behavior of the Affluent and the 1986 Tax Reform Act”. National Tax Association Proceedings, Eighty-Seventh Annual Conference, pp. 70–76.

Auten, G., Carroll, R. (1999). “The Effect of Income Taxes on Household Income”. Review of Economics and Statistics 81, 681–693.

Ballard, C.L., Fullerton, D. (1992). “Distortionary Taxes and the Provision of Public Goods”. Journal of Economic Perspectives 6 (3), 117–131.

Bankman, J., Griffith, T. (1987). “Social Welfare and the Rate Structure: A New Look at Progressive Taxation”. California Law Review 75, 1905–1967.

Barnett, W.A. (1979). “The Joint Allocation of Leisure and Goods Expenditure”. Econometrica 47, 539–563.

Barro, R.J. (1974). “Are Government Bonds Net Wealth?”. Journal of Political Economy 82, 1095–1117.

Barro, R.J. (1979). “On the Determination of the Public Debt”. Journal of Political Economy 87, 940–971.

Becker, G.S. (1968). “Crime and Punishment: An Economic Approach”. Journal of Political Economy 76, 169–217.

Becker, G.S. (1974). “A Theory of Social Interactions”. Journal of Political Economy 82, 1063–1093.

Becker, G.S. (1991). A Treatise on the Family. Harvard University Press, Cambridge, Mass.

Behrman, J.R. (1997). “Intrahousehold Distribution and the Family”. In: Rosenzweig, M.R., Stark, O. (Eds.), Handbook of Population and Family Economics, vol. 1A. Elsevier, Amsterdam, pp. 125–187.

Bergstrom, T.C. (1997). “A Survey of Theories of the Family”. In: Rosenzweig, M.R., Stark, O. (Eds.), Handbook of Population and Family Economics, vol. 1A. Elsevier, Amsterdam, pp. 21–79.

Bernheim, B.D. (2002). “Taxation and Saving”. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 3. Elsevier, Amsterdam, pp. 1173–1249.

Bernheim, B.D., Shleifer, A., Summers, L.H. (1985). “The Strategic Bequest Motive”. Journal of Political Economy 93, 1045–1076.

Besley, T., Coate, S. (1995). “The Design of Income Maintenance Programmes”. Review of Economic Studies 62, 187–221.

Bittker, B.I. (1969). “Accounting for Federal ‘Tax Subsidies’ in the National Budget”. National Tax Journal 22, 244–261.

Bittker, B.I. (1979). “Equity, Efficiency, and Income Tax Theory: Do Misallocations Drive Out Inequities?”. San Diego Law Review 16, 735–748.

Blackorby, C., Donaldson, D. (1988). “Cash Versus Kind, Self-Selection, and Efficient Transfers”. American Economic Review 78, 691–700.

Blair, D.H. (1988). “The Primary-Goods Indexation Problem in Rawls’s ‘Theory of Justice’ ”. Theory and Decision 24, 239–252.

Blanchflower, D.G., Oswald, A.J. (1998). “What Makes an Entrepreneur?”. Journal of Labor Economics 16, 26–60.

Blinder, A.S., Gordon, R.H., Wise, D.E. (1980). “Reconsidering the Work Disincentive Effects of Social Security”. National Tax Journal 33, 431–442.

Blum, W.J., Kalven, H. Jr. (1953). The Uneasy Case for Progressive Taxation. University of Chicago Press, Chicago.

Boadway, R., Keen, M. (1993). “Public Goods, Self-Selection and Optimal Income Taxation”. International Economic Review 34, 463–478.

Boadway, R., Marchand, M., Pestieau, P. (1994). “Towards a Theory of the Direct–Indirect Tax Mix”. Journal of Public Economics 55, 71–88.

Boskin, M.J. (1977). “Notes on the Tax Treatment of Human Capital”. In: U.S. Department of the Treasury, Office of Tax Analysis, Conference on Tax Research 1975. U.S. Government Printing Office, Washington, D.C., pp. 185–195.

Boskin, M.J., Sheshinski, E. (1978). “Optimal Redistributive Taxation When Individual Welfare Depends upon Relative Income”. Quarterly Journal of Economics 92, 589–601.

Boskin, M.J., Sheshinski, E. (1983). “Optimal Tax Treatment of the Family: Married Couples”. Journal of Public Economics 20, 281–297.

Bovenberg, A.L., Goulder, L.H. (2002). “Environmental Taxation and Regulation”. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 3. Elsevier, Amsterdam, pp. 1471–1545.

Bradford, D.F. (1980). “The Economics of Tax Policy toward Savings”. In: von Furstenberg, G.M. (Ed.), The Government and Capital Formation. Ballinger, Cambridge, Mass, pp. 11–71.

Bradford, D.F. (1981). “The Incidence and Allocation Effects of a Tax on Corporate Distributions”. Journal of Public Economics 15, 1–22.

Bradford, D.F. (1986). Untangling the Income Tax. Harvard University Press, Cambridge, Mass.

Bradford, D.F. (1995). “Fixing Realization Accounting: Symmetry, Consistency and Correctness in the Taxation of Financial Instruments”. Tax Law Review 50, 731–785.

Bradford, D.F. (1996a). “Consumption Taxes: Some Fundamental Transition Issues”. In: Boskin, M.J. (Ed.), Frontiers of Tax Reform. Hoover Institution Press, Stanford, pp. 123–150.

Bradford, D.F. (1996b). Fundamental Issues in Consumption Taxation. American Enterprise Institute Press, Washington, D.C.

Bradford, D.F., U.S. Treasury Tax Policy Staff (1984). Blueprints for Basic Tax Reform, 2nd edn. Tax Analysts, Arlington, Va.

Brennan, G. (1976a). "The Distributional Implications of Public Goods". Econometrica 44, 391–399.

Brennan, G. (1976b). "Public Goods and Income Distribution: A Rejoinder to the Aaron-McGuire Reply". Econometrica 44, 405–407.

Brett, C. (1998). "Who Should Be on Workfare? The Use of Work Requirements as Part of an Optimal Tax Mix". Oxford Economic Papers 50, 607–622.

Brito, D.L., Oakland, W.H. (1977). "Some Properties of the Optimal Income Tax". International Economic Review 18, 407–423.

Brito, D.L., Sheshinski, E., Intriligator, M.D. (1991). "Externalities and Compulsory Vaccinations". Journal of Public Economics 45, 69–90.

Broome, J. (1984). "Uncertainty and Fairness". Economic Journal 94, 624–632.

Brown, J.R., Mitchell, O.S., Poterba, J.M. (2002). "Mortality Risk, Inflation Risk, and Annuity Products". In: Mitchell, O.S., et al. (Eds.), Innovations in Retirement Financing. University of Pennsylvania Press, Philadelphia, pp. 175–197.

Bruce, N., Waldman, M. (1991). "Transfers in Kind: Why They Can Be Efficient and Nonpaternalistic". American Economic Review 81, 1345–1351.

Buchanan, J.M. (1975). "The Samaritan's Dilemma". In: Phelps, E. (Ed.), Altruism, Morality, and Economic Theory. Russell Sage Foundation, New York, pp. 71–85.

Bulow, J.I., Summers, L.H. (1984). "The Taxation of Risky Assets". Journal of Political Economy 92, 20–39.

Carruth, A.A. (1982). "On the Role of the Production and Consumption Assumptions for Optimum Taxation". Journal of Public Economics 17, 145–155.

Casler, S.D., Rafiqui, A. (1993). "Evaluating Fuel Tax Equity: Direct and Indirect Distributional Effects". National Tax Journal 46, 197–205.

Chamley, C. (1986). "Optimal Taxation of Capital Income in General Equilibrium with Infinite Lives". Econometrica 54, 607–622.

Christiansen, V. (1981). "Evaluation of Public Projects under Optimal Taxation". Review of Economic Studies 48, 447–457.

Clark, S., Hemming, R., Ulph, D. (1981). "On Indices for the Measurement of Poverty". Economic Journal 91, 515–526.

Coase, R.H. (1960). "The Problem of Social Cost". Journal of Law & Economics 3, 1–44.

Coate, S. (1995). "Altruism, the Samaritan's Dilemma, and Government Transfer Policy". American Economic Review 85, 46–57.

Cochrane, J.H. (1991). "A Simple Test of Consumption Insurance". Journal of Political Economy 99, 957–976.

Corlett, W.J., Hague, D.C. (1953). "Complementarity and the Excess Burden of Taxation". Review of Economic Studies 21, 21–30.

Coronado, J.L., Fullerton, D., Glass, T. (1999). "Distributional Impacts of Proposed Changes to the Social Security System". In: Poterba, J.M. (Ed.), Tax Policy and the Economy, vol. 13. MIT Press, Cambridge, Mass., pp. 149–186.

Cowell, F.A. (1990). Cheating the Government: The Economics of Evasion. MIT Press, Cambridge, Mass.

Cowell, F.A. (1995). Measuring Inequality, 2nd edn. Prentice Hall, London.

Cox, D. (1987). "Motives for Private Income Transfers". Journal of Political Economy 95, 508–546.

Cox, D. (1990). "Intergenerational Transfers and Liquidity Constraints". Quarterly Journal of Economics 105, 187–217.

Cremer, H., Dellis, A., Pestieau, P. (2003). "Family Size and Optimal Income Taxation". Journal of Population Economics 16, 37–54.

Cremer, H., Gahvari, F., Ladoux, N. (1998). "Externalities and Optimal Taxation". Journal of Public Economics 70, 343–364.

Dahan, M., Strawczynski, M. (2000). "Optimal Income Taxation: An Example with a U-Shaped Pattern of Optimal Marginal Tax Rates: Comment". American Economic Review 90, 681–686.

Dahan, M., Strawczynski, M. (2004). "The Optimal Asymptotic Income Tax Rate", Discussion Paper No. 2004.15, Bank of Israel Discussion Paper Series. Research Department, Bank of Israel, Jerusalem.

Dalton, H. (1920). "The Measurement of the Inequality of Incomes". Economic Journal 30, 348–361.

Dasgupta, P., Hammond, P. (1980). "Fully Progressive Taxation". Journal of Public Economics 13, 141–154.

Dasgupta, P., Stiglitz, J. (1972). "On Optimal Taxation and Public Production". Review of Economic Studies 39, 87–103.

Davies, J., Whalley, J. (1991). "Taxes and Capital Formation: How Important Is Human Capital?". In: Bernheim, B.D., Shoven, J.B. (Eds.), National Saving and Economic Performance. University of Chicago Press, Chicago, pp. 163–200.

Deaton, A.S. (1979). "Optimally Uniform Commodity Taxes". Economics Letters 2, 357–361.

Deaton, A.S., Muellbauer, J. (1986). "On Measuring Child Costs: With Applications to Poor Countries". Journal of Political Economy 94, 720–744.

Desai, M.A., Hines, J.R. Jr. (2003). "Evaluating International Tax Reform". National Tax Journal 56, 487–502.

Diamond, P.A. (1967). "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment". Journal of Political Economy 75, 765–766.

Diamond, P.A. (1968). "Negative Taxes and the Poverty Problem—A Review Article". National Tax Journal 21, 288–303.

Diamond, P.A. (1975). "A Many-Person Ramsey Tax Rule". Journal of Public Economics 4, 335–342.

Diamond, P.A. (1980). "Income Taxation with Fixed Hours of Work". Journal of Public Economics 13, 101–110.

Diamond, P.A. (1998). "Optimal Income Taxation: An Example with a U-Shaped Pattern of Optimal Marginal Tax Rates". American Economic Review 88, 83–95.

Diamond, P.A. (2002). Social Security Reform. Oxford University Press, Oxford.

Diamond, P.A. (2003). Taxation, Incomplete Markets, and Social Security: The 2000 Munich Lectures. MIT Press, Cambridge, Mass.

Diamond, P.A. (2004). "Social Security". American Economic Review 94, 1–24.

Diamond, P.A., Mirrlees, J.A. (1971). "Optimal Taxation and Public Production II: Tax Rules". American Economic Review 61, 261–278.

Diamond, P.A., Mirrlees, J.A. (1986). "Payroll-Tax Financed Social Insurance with Variable Retirement". Scandinavian Journal of Economics 88, 25–50.

Dickert, S., Houser, S., Scholz, J.K. (1994). "Taxes and the Poor: A Microsimulation Study of Implicit and Explicit Taxes". National Tax Journal 47, 621–638.

Dixit, A.K. (1975). "Welfare Effects of Tax and Price Changes". Journal of Public Economics 4, 103–123.

Dixit, A.K. (1985). "Tax Policy in Open Economies". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 1. North-Holland, Amsterdam, pp. 313–374.

Domar, E.D., Musgrave, R.A. (1944). "Proportional Income Taxation and Risk-Taking". Quarterly Journal of Economics 58, 388–423.

Dominitz, J., Manski, C.F., Heinz, J. (2003). " 'Will Social Security Be There for You?': How Americans Perceive Their Benefits", Working Paper No. 9798. National Bureau of Economic Research, Cambridge, Mass.

Drèze, J., Stern, N.H. (1987). "The Theory of Cost–Benefit Analysis". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 2. North-Holland, Amsterdam, pp. 909–989.

Duesenberry, J.S. (1949). Income, Saving, and the Theory of Consumer Behavior. Harvard University Press, Cambridge, Mass.

Easterlin, R.A. (1973). "Does Money Buy Happiness?". Public Interest 30, 3–10.

Easterlin, R.A. (1974). "Does Economic Growth Improve the Human Lot? Some Empirical Evidence". In: David, P.A., Reder, M.W. (Eds.), Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz. Academic Press, New York, pp. 89–125.

Easterlin, R.A. (2001). "Income and Happiness: Towards a Unified Theory". Economic Journal 111, 465–484.

Eaton, J., Rosen, H.S. (1980a). "Labor Supply, Uncertainty, and Efficient Taxation". Journal of Public Economics 14, 365–374.

Eaton, J., Rosen, H.S. (1980b). "Optimal Redistributive Taxation and Uncertainty". Quarterly Journal of Economics 95, 357–364.

Eaton, J., Rosen, H.S. (1980c). “Taxation, Human Capital, and Uncertainty”. American Economic Review 70, 705–715.

Ebert, U. (1992). “A Reexamination of the Optimal Nonlinear Income Tax”. Journal of Public Economics 49, 47–73.

Edgeworth, F.Y. (1897). “The Pure Theory of Taxation”. Economic Journal 7, 46–70, 226–238, 550–571.

Feldstein, M.S. (1972). “Equity and Efficiency in Public Sector Pricing: The Optimal Two-Part Tariff”. Quarterly Journal of Economics 86, 175–187.

Feldstein, M.S. (1973). “On the Optimal Progressivity of the Income Tax”. Journal of Public Economics 2, 357–376.

Feldstein, M.S. (1974). “Distributional Preferences in Public Expenditure Analysis”. In: Hochman, H.M., Peterson, G.E. (Eds.), Redistribution through Public Choice. Columbia University Press, New York, pp. 136–161.

Feldstein, M.S. (1976). “On the Theory of Tax Reform”. Journal of Public Economics 6, 77–104.

Feldstein, M.S. (1978). “The Welfare Cost of Capital Income Taxation”. Journal of Political Economy 86 (2, pt2), S29–S51.

Feldstein, M.S. (1983). Inflation, Tax Rules, and Capital Formation. University of Chicago Press, Chicago.

Feldstein, M.S. (1995). “The Effect of Marginal Tax Rates on Taxable Income: A Panel Study of the 1986 Tax Reform Act”. Journal of Political Economy 103, 551–572.

Feldstein, M.S. (1999). “Tax Avoidance and the Deadweight Loss of the Income Tax”. Review of Economics and Statistics 81, 674–680.

Feldstein, M.S. (2005). “Rethinking Social Insurance”. American Economic Review 95, 1–24.

Feldstein, M.S., Feenberg, D. (1996). “The Effect of Increased Tax Rates on Taxable Income and Economic Efficiency: A Preliminary Analysis of the 1993 Tax Rate Increases”. In: Poterba, J.M. (Ed.), Tax Policy and the Economy, vol. 10. MIT Press, Cambridge, Mass., pp. 89–117.

Feldstein, M.S., Liebman, J.B. (Eds.) (2002a). The Distributional Aspects of Social Security and Social Security Reform. University of Chicago Press, Chicago.

Feldstein, M.S., Liebman, J.B. (2002b). “Social Security”. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. Elsevier, Amsterdam, pp. 2245–2324.

Feldstein, M.S., Samwick, A. (1992). “Social Security Rules and Marginal Tax Rates”. National Tax Journal 45, 1–22.

Frank, R.H. (1984a). “Are Workers Paid Their Marginal Products?” American Economic Review 74, 549–571.

Frank, R.H. (1984b). “Interdependent Preferences and the Competitive Wage Structure”. RAND Journal of Economics 15, 510–520.

Frank, R.H. (1985). Choosing the Right Pond: Human Behavior and the Quest for Status. Oxford University Press, New York.

Frank, R.H. (1999). Luxury Fever: Money and Happiness in an Era of Excess. Princeton University Press, Princeton.

Fullerton, D. (1991). “Reconciling Recent Estimates of the Marginal Welfare Cost of Taxation”. American Economic Review 81, 302–308.

Fullerton, D., Rogers, D.L. (1993). Who Bears the Lifetime Tax Burden? Brookings Institution, Washington, D.C.

Giannarelli, L., Steuerle, E. (1995). “The Twice-Poverty Trap: Tax Rates Faced by AFDC Recipients”, Working Paper. The Urban Institute, Washington, D.C.

Gibbard, A. (1979). “Disparate Goods and Rawls’ Difference Principle: A Social Choice Theoretic Treatment”. Theory and Decision 11, 267–288.

Giertz, S. (2004). “Recent Literature on Taxable-Income Elasticities”, Technical Paper Series No. 2004-16. Congressional Budget Office, Washington, D.C.

Gokhale, J., Kotlikoff, L.J., Sluchynsky, A. (2002). “Does It Pay to Work?”, Working Paper No. 9096. National Bureau of Economic Research, Cambridge, Mass.

Golosov, M., Tsyvinski, A., Werning, I. (2007). "New Dynamic Public Finance: A User's Guide". In: Acemoglu, D., Rogoff, K., Woodford, M. (Eds.), NBER Macroeconomics Annual 2006. MIT Press, Cambridge, Mass. (in press).

Goolsbee, A. (1998). "Taxes, Organizational Form, and the Deadweight Loss of the Corporate Income Tax". Journal of Public Economics 69, 143–152.

Goolsbee, A. (2000). "It's Not About the Money: Why Natural Experiments Don't Work on the Rich". In: Slemrod, J.B. (Ed.), Does Atlas Shrug?: The Economic Consequences of Taxing the Rich. Russell Sage Foundation, New York, pp. 141–158.

Goolsbee, A. (2004). "The Impact of the Corporate Income Tax: Evidence from State Organizational Form Data". Journal of Public Economics 88, 2283–2299.

Gordon, R.H. (1985). "Taxation of Corporate Capital Income: Tax Revenues Versus Tax Distortions". Quarterly Journal of Economics 100, 1–27.

Gordon, R.H., Hines, J.R. Jr. (2002). "International Taxation". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. Elsevier, Amsterdam, pp. 1935–1999.

Gordon, R.H., Li, W. (2005). "Puzzling Tax Structures in Developing Countries: A Comparison of Two Alternative Explanations", Working Paper No. 11661. National Bureau of Economic Research, Cambridge, Mass.

Gordon, R.H., MacKie-Mason, J.K. (1994). "Tax Distortions to the Choice of Organizational Form". Journal of Public Economics 55, 279–306.

Gordon, R.H., Slemrod, J. (1988). "Do We Collect Any Revenue from Taxing Capital Income?". In: Summers, L.H. (Ed.), Tax Policy and the Economy, vol. 2. MIT Press, Cambridge, Mass., pp. 89–130.

Gordon, R.H., Kalambokidis, L., Slemrod, J. (2004). "Do We *Now* Collect Any Revenue from Taxing Capital Income?". Journal of Public Economics 88, 981–1009.

Gordon, R.H., Kalambokidis, L., Rohaly, J., Slemrod, J. (2004). "Toward a Consumption Tax, and Beyond". American Economic Review (AEA Papers and Proceedings) 94 (2), 161–165.

Goulder, L.H. (2002). Environmental Policy Making in Economies with Prior Tax Distortions. Edward Elgar, Cheltenham, UK.

Graham, J.R. (2003). "Taxes and Corporate Finance: A Review". Review of Financial Studies 16, 1075–1129.

Gravelle, J.G. (1989). "Differential Taxation of Capital Income: Another Look at the 1986 Tax Reform Act". National Tax Journal 42, 441–463.

Gravelle, J.G. (1994). The Economic Effects of Taxing Capital Income. MIT Press, Cambridge, Mass.

Gravelle, J.G., Kotlikoff, L.J. (1989). "The Incidence and Efficiency Costs of Corporate Taxation When Corporate and Noncorporate Firms Produce the Same Good". Journal of Political Economy 97, 749–780.

Griffith, T.D. (1989). "Theories of Personal Deductions in the Income Tax". Hastings Law Journal 40, 343–395.

Gronau, R. (1988). "Consumption Technology and the Intrafamily Distribution of Resources: Adult Equivalence Scales Reexamined". Journal of Political Economy 96, 1183–1205.

Groves, H.M. (1974). Tax Philosophers: Two Hundred Years of Thought in Great Britain and the United States. University of Wisconsin Press, Madison.

Gruber, J., Mullainathan, S. (2005). "Do Cigarette Taxes Make Smokers Happier". Advances in Economic Analysis & Policy 5 (1), article 4.

Gruber, J., Saez, E. (2002). "The Elasticity of Taxable Income: Evidence and Implications". Journal of Public Economics 84, 1–32.

Gruber, J., Yelowitz, A. (1999). "Public Health Insurance and Private Savings". Journal of Political Economy 107, 1249–1274.

Guyton, J.L., O'Hare, J.F., Stavrianos, M.P., Toder, E.J. (2003). "Estimating the Compliance Cost of the U.S. Individual Income Tax". National Tax Journal 56, 673–688.

Halperin, D., Steuerle, E. (1988). "Indexing the Tax System for Inflation". In: Aaron, H.J., Galper, H., Pechman, J.A. (Eds.), Uneasy Compromise: Problems of a Hybrid Income-Consumption Tax. Brookings Institution, Washington, D.C., pp. 347–380.

Hamilton, B.W. (1976). "Capitalization of Intrajurisdictional Differences in Local Tax Prices". American Economic Review 66, 743–753.

Hamilton, J.H. (1987). “Optimal Wage and Income Taxation with Wage Uncertainty”. International Economic Review 28, 373–388.

Hammond, P. (1983). “Ex-post Optimality as a Dynamically Consistent Objective for Collective Choice under Uncertainty”. In: Pattanaik, P.K., Salles, M. (Eds.), Social Choice and Welfare. Elsevier, New York, pp. 175–205.

Harberger, A.C. (1962). “The Incidence of the Corporation Income Tax”. Journal of Political Economy 70, 215–240.

Harberger, A.C. (1966). “Efficiency Effects of Taxes on Income from Capital”. In: Krzyzaniak, M. (Ed.), Effects of Corporation Income Tax. Wayne State University Press, Detroit, pp. 107–117.

Hare, R.M. (1973). “Rawls’ Theory of Justice—I”. Philosophical Quarterly 23, 144–155.

Harsanyi, J.C. (1953). “Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking”. Journal of Political Economy 61, 434–435.

Harsanyi, J.C. (1955). “Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility”. Journal of Political Economy 63, 309–321.

Harsanyi, J.C. (1975). “Nonlinear Social Welfare Functions: Do Welfare Economists Have a Special Exemption from Bayesian Rationality?”. Theory and Decision 6, 311–332.

Heckman, J.J. (1976). “A Life-Cycle Model of Earnings, Learning, and Consumption”. Journal of Political Economy 84 (4, pt 2), S11–S44.

Heckman, J.J., Lochner, L., Taber, C. (1999). “Human Capital Formation and General Equilibrium Treatment Effects: A Study of Tax and Tuition Policy”. Fiscal Studies 20, 25–40.

Hepner, M., Reed, W.R. (2004). “The Effect of Welfare on Work and Marriage: A View from the States”. Cato Journal 24, 349–370.

Hines, J.R. Jr. (2000). “What Is Benefit Taxation?”. Journal of Public Economics 75, 483–492.

Hochman, H.M., Rodgers, J.D. (1969). “Pareto Optimal Redistribution”. American Economic Review 59, 542–557.

Hoff, K., Lyon, A.B. (1995). “Non-Leaky Buckets: Optimal Redistributive Taxation and Agency Costs”. Journal of Public Economics 58, 365–390.

Holtz-Eakin, D., Joulfaian, D., Rosen, H.S. (1993). “The Carnegie Conjecture: Some Empirical Evidence”. Quarterly Journal of Economics 108, 413–435.

Holtz-Eakin, D., Joulfaian, D., Rosen, H.S. (1994a). “Entrepreneurial Decisions and Liquidity Constraints”. RAND Journal of Economics 25, 334–347.

Holtz-Eakin, D., Joulfaian, D., Rosen, H.S. (1994b). “Sticking It Out: Entrepreneurial Survival and Liquidity Constraints”. Journal of Political Economy 102, 53–75.

Hubbard, R.G., Judd, K.L. (1986). “Liquidity Constraints, Fiscal Policy, and Consumption”. Brookings Papers on Economic Activity 1986 (1), 1–50 (with comments by R.E. Hall, 51–53, and L. Summers, 53–57).

Hurst, E., Ziliak, J.P. (2006). “Do Welfare Asset Limits Affect Household Saving? Evidence from Welfare Reform”. Journal of Human Resources 41, 46–71.

Hylland, A., Zeckhauser, R. (1979). “Distributional Objectives Should Affect Taxes But Not Program Choice or Design”. Scandinavian Journal of Economics 81, 264–284.

Imbens, G.W., Rubin, D.B., Sacerdote, B.I. (2001). “Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players”. American Economic Review 91, 778–794.

Inman, R.P., Rubinfeld, D.L. (1996). “Designing Tax Policy in Federalist Economies: An Overview”. Journal of Public Economics 60 (3), 307–334.

Internal Revenue Service (1996). “Federal Tax Compliance Research: Individual Income Tax Gap Estimates for 1985, 1988, and 1992”. Publication 1415. Internal Revenue Service, Washington, D.C.

Internal Revenue Service (2005). “Tax Gap Facts and Figures”. Internal Revenue Service, Washington, D.C. http://www.irs.gov/pub/irs-utl/tax_gap_facts-figures.pdf.

Iorwerth, A.A., Whalley, J. (2002). “Efficiency Considerations and the Exemption of Food from Sales and Value Added Taxes”. Canadian Journal of Economics 35, 166–182.

Jacobs, B. (2005). “Optimal Income Taxation with Endogenous Human Capital”. Journal of Public Economic Theory 7, 295–315.

Jakobsson, U. (1976). “On the Measurement of the Degree of Progression”. Journal of Public Economics 5, 161–168.

Joint Economic Committee (1999). The Economics of the Estate Tax. Government Printing Office, Washington, D.C.

Jorgenson, D.W., Fraumeni, B.M. (1989). “The Accumulation of Human and Nonhuman Capital, 1948–84”. In: Lipsey, R.E., Tice, H.S. (Eds.), The Measurement of Saving, Investment, and Wealth. University of Chicago Press, Chicago, pp. 227–282.

Joulfaian, D., Wilhelm, M.O. (1994). “Inheritance and Labor Supply”. Journal of Human Resources 29, 1205–1234.

Judd, K.L. (1985). “Redistributive Taxation in a Simple Perfect Foresight Model”. Journal of Public Economics 28, 59–83.

Kakwani, N. (1977). “Applications of Lorenz Curves in Economic Analysis”. Econometrica 45, 719–728.

Kanbur, R., Tuomala, M. (1994). “Inherent Inequality and the Optimal Graduation of Marginal Tax Rates”. Scandinavian Journal of Economics 96, 275–282.

Kaplow, L. (1989). “Horizontal Equity: Measures in Search of a Principle”. National Tax Journal 42, 139–154.

Kaplow, L. (1990). “Optimal Taxation with Costly Enforcement and Evasion”. Journal of Public Economics 43, 221–236.

Kaplow, L. (1992). “Income Tax Deductions for Losses as Insurance”. American Economic Review 82, 1013–1017.

Kaplow, L. (1994). “Taxation and Risk Taking: A General Equilibrium Perspective”. National Tax Journal 47, 789–798.

Kaplow, L. (1995). “A Fundamental Objection to Tax Equity Norms: A Call for Utilitarianism”. National Tax Journal 48, 497–514.

Kaplow, L. (1996a). “On the Divergence Between ‘Ideal’ and Conventional Income-Tax Treatment of Human Capital”. American Economic Review (AEA Papers and Proceedings) 86 (2), 347–352.

Kaplow, L. (1996b). “Optimal Distribution and the Family”. Scandinavian Journal of Economics 98, 75–92.

Kaplow, L. (1996c). “The Optimal Supply of Public Goods and the Distortionary Cost of Taxation”. National Tax Journal 49, 513–533.

Kaplow, L. (1998). “Accuracy, Complexity, and the Income Tax”. Journal of Law, Economics, & Organization 14, 61–83.

Kaplow, L. (2001). “A Framework for Assessing Estate and Gift Taxation”. In: Gale, W.G., Hines, J.R. Jr., Slemrod, J. (Eds.), Rethinking Estate and Gift Taxation. Brookings Institution, Washington, D.C., pp. 164–215.

Kaplow, L. (2004). “On the (Ir)relevance of Distribution and Labor Supply Distortion to Government Policy”. Journal of Economic Perspectives 18 (4), 159–175.

Kaplow, L. (2005). “Why Measure Inequality?”. Journal of Economic Inequality 3, 65–79.

Kaplow, L. (2006a). “Myopia and the Effects of Social Security and Capital Taxation on Labor Supply”, Working Paper No. 12452. National Bureau of Economic Research, Cambridge, Mass.

Kaplow, L. (2006b). “On the Undesirability of Commodity Taxation Even When Income Taxation Is Not Optimal”. Journal of Public Economics 90, 1235–1250.

Kaplow, L. (2006c). “Optimal Control of Externalities in the Presence of Income Taxation”, Working Paper No. 12339. National Bureau of Economic Research, Cambridge, Mass.

Kaplow, L. (2006d). “Public Goods and the Distribution of Income”. European Economic Review 50, 1627–1660.

Kaplow, L. (2007a). “Capital Levies and Transition to a Consumption Tax”. In: Auerbach, A.J., Shaviro, D. (Eds.), Institutional Foundations of Public Finance: Economic and Legal Perspectives. Harvard University Press, Cambridge, Mass. (in press).

Kaplow, L. (2007b). “Optimal Income Transfers”. International Tax and Public Finance 14, 295–325.

Kaplow, L. (in press). The Theory of Taxation and Public Economics. Princeton University Press, Princeton.

Kaplow, L., Shavell, S. (1994). “Why the Legal System Is Less Efficient Than the Income Tax in Redistributing Income”. Journal of Legal Studies 23, 667–681.

Kaplow, L., Shavell, S. (2001). "Any Non-Welfarist Method of Policy Assessment Violates the Pareto Principle". Journal of Political Economy 109, 281–286.

Kaplow, L., Shavell, S. (2002). Fairness Versus Welfare. Harvard University Press, Cambridge, Mass.

Katz, A., Mankiw, G. (1985). "How Should Fringe Benefits Be Taxed?" National Tax Journal 38, 37–46.

Kesselman, J.R. (1989). "Income Tax Evasion: An Intersectoral Analysis". Journal of Public Economics 38, 137–182.

Kesselman, J.R. (1993). "Evasion Effects of Changing the Tax Mix". Economic Record 69, 131–148.

King, M.A. (1974). "Taxation and the Cost of Capital". Review of Economic Studies 41, 21–35.

King, M.A. (1977). Public Policy and the Corporation. Chapman and Hall, London.

King, M.A. (1983). "An Index of Inequality: With Applications to Horizontal Equity and Social Mobility". Econometrica 51, 99–115.

Konishi, H. (1995). "A Pareto-Improving Commodity Tax Reform under a Smooth Nonlinear Income Tax". Journal of Public Economics 56, 413–446.

Kopczuk, W. (2005). "Tax Bases, Tax Rates and the Elasticity of Reported Income". Journal of Public Economics 89, 2093–2119.

Kotlikoff, L.J. (2002). "Generational Policy". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. Elsevier, Amsterdam, pp. 1873–1932.

Kydland, F.E., Prescott, E.C. (1977). "Rules Rather than Discretion: The Inconsistency of Optimal Plans". Journal of Political Economy 85, 473–491.

Laibson, D. (1998). "Life-cycle Consumption and Hyperbolic Discount Functions". European Economic Review 42, 861–871.

Lambert, P.J. (2001). The Distribution and Redistribution of Income, 3rd edn. Manchester University Press, Manchester.

Laroque, G. (2005). "Indirect Taxation Is Superfluous under Separability and Taste Homogeneity: A Simple Proof". Economics Letters 87, 141–144.

Leimer, D.R. (1994). "Cohort Specific Measures of Lifetime Net Social Security Transfers", Working Paper 59. Office of Research and Statistics, Social Security Administration, Washington, D.C.

Liebman, J.B. (2002). "Redistribution in the Current U.S. Social Security System". In: Feldstein, M., Liebman, J.B. (Eds.), The Distributional Aspects of Social Security and Social Security Reform. University of Chicago Press, Chicago, pp. 11–48.

Liebman, J.B. (2003). "Should Taxes Be Based on Lifetime Income? Vickrey Taxation Revisited", Preliminary Draft Manuscript.

Lindahl, E. (1919). Die Gerechtigkeit der Besteuerung: Eine Analyse der Steuerprinzipien auf der Grundlage der Grenznutzentheorie. Gleerup and H. Ohlsson, Lund, ch. 4, "Positive Lösung", translated as "Just Taxation—A Positive Solution". In: Musgrave, R.A., Peacock, A.T. (Eds.), Classics in the Theory of Public Finance. Macmillan, London, 1958, pp. 168–176.

Little, I.M.D. (1957). A Critique of Welfare Economics, 2nd edn. Clarendon Press, Oxford.

Low, H., Maldoom, D. (2004). "Optimal Taxation, Prudence and Risk-Sharing". Journal of Public Economics 88, 443–464.

Mace, B.J. (1991). "Full Insurance in the Presence of Aggregate Uncertainty". Journal of Political Economy 99, 928–956.

MacKie-Mason, J.K., Gordon, R.H. (1997). "How Much Do Taxes Discourage Incorporation?". Journal of Finance 52, 477–505.

Mayshar, J. (1990). "On Measures of Excess Burden and Their Application". Journal of Public Economics 43, 263–289.

Mayshar, J. (1991). "Taxation with Costly Administration". Scandinavian Journal of Economics 93, 75–88.

McLure, C.E. Jr. (1979). Must Corporate Income Be Taxed Twice?. Brookings Institution, Washington, D.C.

Messere, K., de Kam, F., Heady, C. (2003). Tax Policy: Theory and Practice in OECD Countries. Oxford University Press, Oxford.

Michalopoulos, C., Robins, P.K., Card, D. (2005). "When Financial Work Incentives Pay for Themselves: Evidence from a Randomized Social Experiment for Welfare Recipients". Journal of Public Economics 89, 5–29.

Mieszkowski, P. (1972). "The Property Tax: An Excise Tax or a Profits Tax?". Journal of Public Economics 1, 73–96.

Mieszkowski, P., Zodrow, G.R. (1989). "Taxation and the Tiebout Model: The Differential Effects of Head Taxes, Taxes on Land Rents, and Property Taxes". Journal of Economic Literature 27, 1098–1146.

Mirrlees, J.A. (1971). "An Exploration in the Theory of Optimum Income Taxation". Review of Economic Studies 38, 175–208.

Mirrlees, J.A. (1972a). "On Producer Taxation". Review of Economic Studies 39, 105–111.

Mirrlees, J.A. (1972b). "Population Policy and the Taxation of Family Size". Journal of Public Economics 1, 169–198.

Mirrlees, J.A. (1976). "Optimal Tax Theory: A Synthesis". Journal of Public Economics 6, 327–358.

Mirrlees, J.A. (1990). "Taxing Uncertain Incomes". Oxford Economic Papers 42, 34–45.

Mirrlees, J.A. (1994). "Optimal Taxation and Government Finance". In: Quigley, J.M., Smolensky, E. (Eds.), Modern Public Finance. Harvard University Press, Cambridge, Mass., pp. 213–231.

Moffitt, R.A. (1992). "Incentive Effects of the U.S. Welfare System: A Review". Journal of Economic Literature 30, 1–61.

Moffitt, R.A. (2002). "Welfare Programs and Labor Supply". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. Elsevier, Amsterdam, pp. 2393–2430.

Moffitt, R.A., Wilhelm, M.O. (2000). "Taxation and the Labor Supply Decisions of the Affluent". In: Slemrod, J.B. (Ed.), Does Atlas Shrug?: The Economic Consequences of Taxing the Rich. Russell Sage Foundation, New York, pp. 193–234.

Munro, A. (1989). "In-Kind Transfers, Cash Grants and the Supply of Labour". European Economic Review 33, 1597–1604.

Musgrave, R.A. (1959). The Theory of Public Finance. McGraw-Hill, New York.

Musgrave, R.A. (1985). "A Brief History of Fiscal Doctrine". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 1. North-Holland, Amsterdam, pp. 1–59.

Musgrave, R.A. (1990). "Horizontal Equity, Once More". National Tax Journal 43, 113–122.

Musgrave, R.A., Musgrave, P.B. (1973). Public Finance in Theory and Practice. McGraw-Hill, New York.

Musgrave, R.A., Musgrave, P.B. (1989). Public Finance in Theory and Practice, 5th edn. McGraw-Hill, New York.

Musgrave, R.A., Thin, T. (1948). "Income Tax Progression 1929–48". Journal of Political Economy 56, 498–514.

Musgrave, R.A., Case, K.E., Leonard, H. (1974). "The Distribution of Fiscal Burdens and Benefits". Public Finance Quarterly 2, 259–311.

Myerson, R.B. (1981). "Utilitarianism, Egalitarianism, and the Timing Effect in Social Choice Problems". Econometrica 49, 883–897.

Nerlove, M., Razin, A., Sadka, E. (1986). "Some Welfare Theoretic Implications of Endogenous Fertility". International Economic Review 27, 3–31.

Nerlove, M., Razin, A., Sadka, E., von Weizsäcker, R.K. (1993). "Comprehensive Income Taxation, Investments in Human and Physical Capital, and Productivity". Journal of Public Economics 50, 397–406.

Ng, Y.-K. (2000). "The Optimal Size of Public Spending and the Distortionary Cost of Taxation". National Tax Journal 53, 253–272.

Nichols, A.L., Zeckhauser, R.J. (1982). "Targeting Transfers Through Restrictions on Recipients". American Economic Review (AEA Papers and Proceedings) 72 (2), 372–377.

Oates, W.E. (1972). Fiscal Federalism. Harcourt Brace Jovanovich, New York.

Oates, W.E. (1999). "An Essay on Fiscal Federalism". Journal of Economic Literature 37, 1120–1149.

Ordover, J.A., Phelps, E.S. (1979). "The Concept of Optimal Taxation in the Overlapping-Generations Model of Capital and Wealth". Journal of Public Economics 12, 1–26.

Pauly, M.V. (1973). "Income Redistribution as a Local Public Good". Journal of Public Economics 2, 35–58.

Phelps, E.S. (1973). "The Taxation of Wage Income for Economic Justice". Quarterly Journal of Economics 87, 331–354.

Pigou, A.C. (1920). The Economics of Welfare. Macmillan, London.

Pigou, A.C. (1928). A Study in Public Finance. Macmillan, London.

Pirttilä, J., Tuomala, M. (1997). "Income Tax, Commodity Tax and Environmental Policy". International Tax and Public Finance 4, 379–393.

Plotnick, R. (1981). "A Measure of Horizontal Inequity". Review of Economics and Statistics 63, 283–288.

Polinsky, A.M. (1974). "Imperfect Capital Markets, Intertemporal Redistribution, and Progressive Taxation". In: Hochman, H.M., Peterson, G.E. (Eds.), Redistribution through Public Choice. Columbia University Press, New York, pp. 229–258.

Polinsky, A.M., Shavell, S. (1979). "The Optimal Tradeoff between the Probability and Magnitude of Fines". American Economic Review 69, 880–891.

Polinsky, A.M., Shavell, S. (2007). "The Theory of Public Enforcement of Law". In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics, vol. 1. Elsevier, Amsterdam, pp. 403–454.

Pollak, R.A., Wales, T.J. (1979). "Welfare Comparisons and Equivalence Scales". American Economic Review (AEA Papers and Proceedings) 69 (2), 216–221.

Poterba, J.M. (2002). "Taxation, Risk-Taking, and Household Portfolio Behavior". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 3. Elsevier, Amsterdam, pp. 1109–1171.

Poterba, J.M., Summers, L.H. (1985). "The Economic Effects of Dividend Taxation". In: Altman, E., Subrahmanyam, M. (Eds.), Recent Advances in Corporate Finance. Richard D. Irwin, Homewood, Ill., pp. 227–284.

Powers, E. (1998). "Does Means-Testing Welfare Discourage Saving? Evidence from a Change in AFDC Policy in the United States". Journal of Public Economics 68, 33–53.

Ramsey, F.P. (1927). "A Contribution to the Theory of Taxation". Economic Journal 37, 47–61.

Ravallion, M. (1994). Poverty Comparisons. Harwood Academic Publishers, Chur.

Rawls, J. (1971). A Theory of Justice. Harvard University Press, Cambridge, Mass.

Rawls, J. (1982). "Social Unity and Primary Goods". In: Sen, A., Williams, B. (Eds.), Utilitarianism and Beyond. Cambridge University Press, New York, pp. 159–185.

Revesz, R.L., Stavins, R.N. (2007). "Environmental Law". In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics, vol. 1. Elsevier, Amsterdam, pp. 499–589.

Reynolds, M., Smolensky, E. (1977). Public Expenditures, Taxes, and the Distribution of Income: The United States, 1950, 1961, 1970. Academic Press, New York.

Richman, P.B. (1963). Taxation of Foreign Investment Income: An Economic Analysis. Johns Hopkins Press, Baltimore.

Roberts, K. (1984). "The Theoretical Limits to Redistribution". Review of Economic Studies 51, 177–195.

Rosen, H.S. (1976). "A Methodology for Evaluating Tax Reform Proposals". Journal of Public Economics 6, 105–121.

Rosen, S. (1986). "The Theory of Equalizing Differences". In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, vol. 1. North-Holland, Amsterdam, pp. 641–692.

Roth, J.A., Scholz, J.T., Witte, A.D. (Eds.) (1989). Taxpayer Compliance, vol. 1, An Agenda for Research; vol. 2, Social Science Perspectives. University of Pennsylvania Press, Philadelphia.

Rubinfeld, D.L. (1987). "The Economics of the Local Public Sector". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 2. North-Holland, Amsterdam, pp. 571–645.

Ruggles, P. (1990). Drawing the Line: Alternative Poverty Measures and Their Implications for Public Policy. Urban Institute Press, Washington, D.C.

Ruggles, P., O'Higgins, M. (1981). "The Distribution of Public Expenditure among Households in the United States". Review of Income and Wealth 27, 137–164.

Sadka, E. (1976). "On Income Distribution, Incentive Effects and Optimal Income Taxation". Review of Economic Studies 43, 261–267.

Saez, E. (2001). "Using Elasticities to Derive Optimal Income Tax Rates". Review of Economic Studies 68, 205–229.

Saez, E. (2002). "Optimal Income Transfer Programs: Intensive versus Extensive Labor Supply Responses". Quarterly Journal of Economics 117, 1039–1073.

Salanié, B. (2003). The Economics of Taxation. MIT Press, Cambridge, Mass.

Sammartino, F., Toder, E., Maag, E. (2002). "Providing Federal Assistance for Low-Income Families through the Tax System: A Primer", Discussion Paper No. 4. Urban-Brookings Tax Policy Center, Washington, D.C.

Samuelson, P.A. (1954). "The Pure Theory of Public Expenditure". Review of Economics and Statistics 36, 387–389.

Samuelson, P.A. (1956). "Social Indifference Curves". Quarterly Journal of Economics 70, 1–22.

Samuelson, P.A. (1964). "Tax Deductibility of Economic Depreciation to Insure Invariant Valuations". Journal of Political Economy 72, 604–606.

Sandmo, A. (1976). "Optimal Taxation: An Introduction to the Literature". Journal of Public Economics 6, 37–54.

Sarkar, S., Zodrow, G.R. (1993). "Transitional Issues in Moving to a Direct Consumption Tax". National Tax Journal 46, 359–376.

Schroyen, F. (2003). "Redistributive Taxation and the Household: The Case of Individual Filings". Journal of Public Economics 87, 2527–2547.

Schultz, T.P. (1994). "Marital Status and Fertility in the United States: Welfare and Labor Market Effects". Journal of Human Resources 29, 637–669.

Scotchmer, S. (2002). "Local Public Goods and Clubs". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. Elsevier, Amsterdam, pp. 1997–2042.

Seade, J.K. (1977). "On the Shape of Optimal Tax Schedules". Journal of Public Economics 7, 203–235.

Sen, A. (1976). "Poverty: An Ordinal Approach to Measurement". Econometrica 44, 219–231.

Sen, A. (1985). Commodities and Capabilities. North-Holland, Amsterdam.

Sen, A. (1997). On Economic Inequality, enlarged edn. Clarendon Press, Oxford.

Sen, A., Williams, B. (Eds.) (1982). Utilitarianism and Beyond. Cambridge University Press, Cambridge.

Shavell, S. (1981). "A Note on Efficiency vs. Distributional Equity in Legal Rulemaking: Should Distributional Equity Matter Given Optimal Income Taxation?". American Economic Review (AEA Papers and Proceedings) 71 (2), 414–418.

Shavell, S. (1991). "Individual Precautions to Prevent Theft: Private versus Socially Optimal Behavior". International Review of Law and Economics 11, 123–132.

Shavell, S. (2007). "Liability for Accidents". In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics, vol. 1. Elsevier, Amsterdam, pp. 139–182.

Shaviro, D. (2000). When Rules Change: An Economic and Political Analysis of Transition Relief and Retroactivity. University of Chicago Press, Chicago.

Shaviro, D. (2004). "Rethinking Tax Expenditures and Fiscal Language". Tax Law Review 57, 187–231.

Silber, J. (Ed.) (1999). Handbook of Income Inequality Measurement. Kluwer Academic Publishers, Boston.

Simons, H.C. (1938). Personal Income Taxation. University of Chicago Press, Chicago.

Slemrod, J. (1988). "Effect of Taxation with International Capital Mobility". In: Aaron, H.J., Galper, H., Pechman, J.A. (Eds.), Uneasy Compromise: Problems of a Hybrid Income-Consumption Tax. Brookings Institution, Washington, D.C., pp. 115–148.

Slemrod, J. (1990). "Optimal Taxation and Optimal Tax Systems". Journal of Economic Perspectives 4 (1), 157–178.

Slemrod, J. (1996a). "High-Income Families and the Tax Changes of the 1980s: The Anatomy of Behavioral Response". In: Feldstein, M., Poterba, J.M. (Eds.), Empirical Foundations of Household Taxation. University of Chicago Press, Chicago, pp. 169–189.

Slemrod, J. (1996b). "Which Is the Simplest Tax System of Them All?". In: Aaron, H.J., Gale, W.G. (Eds.), The Economic Effects of Fundamental Tax Reform. Brookings Institution Press, Washington, D.C., pp. 355–391.

Slemrod, J. (1998). "Methodological Issues in Measuring and Interpreting Taxable Income Elasticities". National Tax Journal 51, 773–788.

Slemrod, J. (2001). "A General Model of the Behavioral Response to Taxation". International Tax and Public Finance 8, 119–128.

Slemrod, J., Kopczuk, W. (2002). "The Optimal Elasticity of Taxable Income". Journal of Public Economics 84, 91–112.

Slemrod, J., Yitzhaki, S. (1987). “The Optimal Size of a Tax Collection Agency”. Scandinavian Journal of Economics 89, 183–192.

Slemrod, J., Yitzhaki, S. (1996). “The Costs of Taxation and the Marginal Efficiency Cost of Funds”. International Monetary Fund Staff Papers 43, 172–198.

Slemrod, J., Yitzhaki, S. (2001). “Integrating Expenditure and Tax Decisions: The Marginal Costs of Funds and the Marginal Benefit of Projects”. National Tax Journal 54, 189–201.

Slemrod, J., Yitzhaki, S. (2002). “Tax Avoidance, Evasion, and Administration”. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 3. Elsevier, Amsterdam, pp. 1423–1470.

Slemrod, J., Yitzhaki, S., Mayshar, J., Lundholm, M. (1994). “The Optimal Two-Bracket Linear Income Tax”. Journal of Public Economics 53, 269–290.

Snow, A., Warren, R.S. Jr. (1983). “Tax Progression in Lindahl Equilibrium”. Economics Letters 12, 319–326.

Stern, N.H. (1976). “On the Specification of Models of Optimum Income Taxation”. Journal of Public Economics 6, 123–162.

Stern, N.H. (1982). “Optimum Taxation with Errors in Administration”. Journal of Public Economics 17, 181–211.

Stiglitz, J.E. (1973). “Taxation, Corporate Financial Policy, and the Cost of Capital”. Journal of Public Economics 2, 1–34.

Stiglitz, J.E. (1982a). “Self-Selection and Pareto Efficient Taxation”. Journal of Public Economics 17, 213–240.

Stiglitz, J.E. (1982b). “Utilitarianism and Horizontal Equity: The Case for Random Taxation”. Journal of Public Economics 18, 1–33.

Stiglitz, J.E. (1985a). “The General Theory of Tax Avoidance”. National Tax Journal 38, 325–337.

Stiglitz, J.E. (1985b). “Inequality and Capital Taxation”, IMSSS Technical Report No. 457. Stanford University, Stanford.

Stiglitz, J.E. (1987). “Pareto Efficient and Optimal Taxation and the New New Welfare Economics”. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 2. North-Holland, Amsterdam, pp. 991–1042.

Stiglitz, J.E. (2000). Economics of the Public Sector, 3rd edn. W.W. Norton, New York.

Stiglitz, J.E., Dasgupta, P. (1971). “Differential Taxation, Public Goods, and Economic Efficiency”. Review of Economic Studies 38, 151–174.

Strawczynski, M. (1998). “Social Insurance and the Optimum Piecewise Linear Income Tax”. Journal of Public Economics 69, 371–388.

Strotz, R.H. (1958). “How Income Ought to Be Distributed: A Paradox in Distributive Ethics”. Journal of Political Economy 66, 189–205.

Suits, D.B. (1977). “Measurement of Tax Progressivity”. American Economic Review 67, 747–752.

Surrey, S.S. (1973). Pathways to Tax Reform: The Concept of Tax Expenditures. Harvard University Press, Cambridge, Mass.

Surrey, S.S., McDaniel, P.R. (1985). Tax Expenditures. Harvard University Press, Cambridge, Mass.

Taubman, P. (1996). “The Roles of the Family in the Formation of Offsprings’ Earnings and Income Capacity”. In: Menchik, P.L. (Ed.), Household and Family Economics. Kluwer Academic Publishers, Boston, pp. 5–45.

Thomson, P. (1965). “The Property Tax and the Rate of Interest”. In: Benson, G.C.S., Benson, S., McClelland, H., Thomson, P. (Eds.), The American Property Tax. Lincoln School of Public Finance, Claremont, pp. 111–198.

Tiebout, C.M. (1956). “A Pure Theory of Local Expenditures”. Journal of Political Economy 64, 416–424.

Trostel, P.A. (1993). “The Effect of Taxation on Human Capital”. Journal of Political Economy 101, 327–350.

Tuomala, M. (1990). Optimal Income Tax and Redistribution. Clarendon Press, Oxford.

U.S. Department of the Treasury (1992). Report of the Department of the Treasury on Integration of the Individual and Corporate Tax Systems: Taxing Business Income Once. Department of the Treasury, Washington, D.C.

Varian, H.R. (1980). “Redistributive Taxation as Social Insurance”. Journal of Public Economics 14, 49–68.

Veblen, T. (1899). The Theory of the Leisure Class. Macmillan, New York.

Veenhoven, R. (1991). "Is Happiness Relative?". Social Indicators Research 24, 1–34.

Vickrey, W. (1939). "Averaging of Income for Income-Tax Purposes". Journal of Political Economy 47, 379–397.

Vickrey, W. (1945). "Measuring Marginal Utility by Reactions to Risk". Econometrica 13, 319–333.

Weisbach, D.A. (2000). "An Efficiency Analysis of Line Drawing in the Tax Law". Journal of Legal Studies 29, 71–97.

Weisbach, D.A., Nussim, J. (2004). "The Integration of Tax and Spending Programs". Yale Law Journal 113, 955–1028.

Weisbrod, B.A. (1968). "Income Redistribution Effects and Benefit-Cost Analysis". In: Chase, S.B. Jr. (Ed.), Problems in Public Expenditure Analysis. Brookings Institution, Washington, D.C., pp. 177–209.

Weitzman, M.L. (1974). "Prices vs. Quantities". Review of Economic Studies 41, 477–491.

West, S.E. (2004). "Distributional Effects of Alternative Vehicle Pollution Control Policies". Journal of Public Economics 88, 735–757.

West, S.E., Williams, R.C. III (2007). "Optimal Taxation and Cross-Price Effects on Labor Supply: Estimates of the Optimal Gas Tax". Journal of Public Economics 91, 593–617.

Whittington, L.A., Alm, J., Peters, H.E. (1990). "Fertility and the Personal Exemption: Implicit Pronatalist Policy in the United States". American Economic Review 80, 545–556.

Wildasin, D.E. (1986). Urban Public Finance. Harwood Academic Publishers, Chur.

Wilson, J.D. (1989). "On the Optimal Tax Base for Commodity Taxation". American Economic Review 79, 1196–1206.

Wilson, J.D. (1999). "Theories of Tax Competition". National Tax Journal 52, 269–304.

Wilson, P., Cline, R. (1994). "State Welfare Reform: Integrating Tax Credits and Income Transfers". National Tax Journal 47, 655–676.

Yitzhaki, S. (1974). "A Note on Income Tax Evasion: A Theoretical Analysis". Journal of Public Economics 3, 201–202.

Yitzhaki, S. (1979). "A Note on Optimal Taxation and Administrative Costs". American Economic Review 69, 475–480.

Yitzhaki, S. (1987). "On the Excess Burden of Tax Evasion". Public Finance Quarterly 15, 123–137.

Zeckhauser, R.J. (1981). Using the Wrong Tool: The Pursuit of Redistribution through Regulation. Chamber of Commerce of the United States, Washington, D.C.

Zodrow, G.R. (1991). "On the 'Traditional' and 'New' Views of Dividend Taxation". National Tax Journal 44, 497–509.

This page intentionally left blank

Chapter 11

INTERNATIONAL LAW

ALAN O. SYKES

School of Law, Stanford University

Contents

Handbook of Law and Economics, Volume 1
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)01011-0

Abstract

This chapter reviews and synthesizes the work of economists and law and economics scholars in the field of public international law. The bulk of that work has been in the area of international trade, but many of the ideas in the trade literature have implications for other subfields. Recent years have seen a significant increase in research on other topics as well. The paper begins with a general framework for thinking about the positive and normative economics of public international law, and then proceeds to a treatment of specific topics including customary law, strategic alliances and the laws of war, international trade, international investment, international antitrust, human rights law, conflicts of law, and the international commons (fisheries).

Keywords

international law, trade, treaties, investment, international antitrust, human rights

JEL classification: F10, K00, K19

1. Introduction

International law has been recognized as a distinct field of study within the legal academy for well over a century, but economically-oriented scholars have paid it relatively little attention. Dunoff and Trachtman (1999). By far the bulk of the law and economics research in international law pertains to the law of international trade. Systematic work on other topics is limited at best, although research in the field generally is accelerating and the subject can properly be considered a growth area.

The relative dearth of prior work, especially formal work, and its concentration on international trade issues, poses a number of challenges for a survey of this sort. An excessive emphasis on international trade will mask the richness and diversity of the field, and obscure rather than illuminate the potential research agenda. Moreover, large segments of the economic literature pertaining to international trade law already receive attention in the three volume *Handbook of International Economics* series, especially in Staiger (1995a), and in a number of other extant and forthcoming volumes focused on WTO issues including Bagwell and Staiger (2002) and Grossman and Helpman (2002). Accordingly, I have limited the treatment of trade issues in this chapter, emphasizing topics that illustrate broader themes for the economic analysis of international law. Much of the chapter will instead be devoted to a general framework for thinking about international law, along with brief discussions of a number of topics outside the realm of trade, with the understanding that an informal treatment is all that the existing literature has to offer on many of them.

Sections 2 and 3 of the chapter provide legal background on the field of international law, followed by a discussion of general economic considerations that cut across a range of possible topics. The analysis will encompass the various possible functions of international law, the challenges involved in realizing gains from international cooperation, the design and function of mechanisms for its enforcement, and the interface between domestic and international law. Sections 4–10 of the chapter consider particular topics, including strategic alliances and laws of war, international trade, international investment, international antitrust, human rights law, conflicts of law, and the international commons (fisheries). I provide only a brief treatment of intellectual property issues (TRIPs) along the way, and devote little attention to environmental issues as well, leaving those subjects to other chapters in this *Handbook* by Menell and Scotchmer and Revesz and Stavins.

As a final, preliminary disclaimer, I have not undertaken to survey and incorporate the vast political science literature on public international law. Much of that literature is excellent, and the "rational choice" literature in particular is often quite close in both spirit and method to the work of economists. I omit attention to it not because of any negative judgment about its quality, but to make the task at hand a manageable one. Readers seeking a window into the rational choice perspective on international relations and institutions might wish to consult Snidal (1996, 2002). Carlsnaes, Risse, and Simmons (2002) provide a broader introduction to modern international relations work in political science.

2. Legal background

The field of international law is conventionally divided into two subfields: "public" international law and "private" international law. "Public" international law refers to the body of law that governs relations between states or countries. "Private" international law refers to the body of law that governs international relations between private citizens or companies. Most private international law relates to international business transactions, and may be subsumed for analytic purposes under other topics such as contract law, corporate law and tax law that are the subject of other chapters in the *Handbook of Law and Economics*. Accordingly, my focus in this chapter is primarily on public international law.

The genesis of public international law necessarily differs from that of domestic law. No international legislature exists to pass the equivalent of domestic statutes, and no international court exists with the power to create a general international common law. Instead, public international law arises only by agreement among states.

Often, agreement is manifest in an instrument known as a *treaty*. A treaty is an agreement executed by duly authorized officials of signatory states, evincing an intention to make it a binding legal obligation. Treaty obligations are themselves governed by an over-arching treaty known as the Vienna Convention on Treaties, which supplies rules for their interpretation and enforcement. Note that the concept of a "treaty" is not necessarily the same in international and domestic law. The U.S. President, for example, has the authority in many areas to bind the United States internationally through "Executive Agreements," sometimes accompanied by formal Congressional approval and sometimes not. These agreements have the same status as treaties under international law, even though they are not treaties under domestic law (which provides that "treaties" must be approved by a two-thirds vote of the U.S. Senate).

Agreement may also become manifest as *customary international law*, which is defined as a "general and consistent practice of states followed by them from a sense of legal obligation."[1] The traditional test for the emergence of customary law thus requires a high degree of consistency in state practice, and a belief that the practice has become a legal obligation. Both of these requirements are imprecise, and scholars often disagree about what practices have achieved the status of customary law. Some rules of customary international law are uncontroversial, however, such as those relating to aspects of diplomatic immunity. It is generally said that states may avoid an obligation to obey customary international law by "opting out" at an early stage of its evolution, but once they have manifested agreement with it through conforming behavior, any subsequent deviation is illegal.

International legal scholars also make reference to the concept of "soft law." Soft law encompasses a range of things, including formal agreements that are understood

[1] American Law Institute, Restatement (Third) of the Foreign Relations Law of the United States §102(2) (1987).

not to be “binding” under international law, as well as agreements that may be “binding” but that are essentially hortatory or aspirational. Examples of each abound—the Cuban Missile Crisis was settled by an informal agreement, for example, while numerous provisions in WTO treaty text encourage but do not require special trade treatment for developing countries.

The enforcement of international law, to the extent that it is successful, occurs in a variety of ways. The closest analog to the coercive enforcement powers often exercised by domestic courts is found in the United Nations. A serious breach of U.N. obligations may result in the authorization of substantial sanctions by the Security Council, or in extreme cases in a resolution authorizing the use of military force against the violator state. Much of international law falls outside the purview of U.N. obligations, however, and thus outside its enforcement mechanism. Some international legal regimes have their own tribunals with the power to adjudicate violations (such as NAFTA and the WTO). The International Court of Justice also has jurisdiction to hear a broad range of disputes. The power to adjudicate disputes may or may not be accompanied by the power to authorize or impose sanctions, however, and the nature of any sanctions may be tightly circumscribed. Formal sanctions for the violation of WTO obligations, for example, are limited to the withdrawal of benefits under WTO agreements. If international law is incorporated into domestic law, as quite often occurs, then the powers of domestic courts can be brought to bear on certain types of violations. This mechanism too has its limitations, as many international legal obligations are never incorporated into domestic law. Further, domestic courts are often limited in their jurisdiction to enforce international obligations that are so incorporated—principles of foreign sovereign immunity, for example, often insulate states from actions against them in foreign courts. Finally, many international legal obligations exist as to which there is no formal enforcement or sanctioning mechanism at all.

Where some enforcement mechanism exists, a further issue arises as to who has *standing* to invoke it. Public international law governs relations among states and, generally speaking, only states have standing to enforce it. Private citizens have no right to pursue most claims under international law even if they have suffered substantial injury due to a violation. An important exception of sorts exists, however, if nations incorporate international law into their domestic legal systems. Private actors may then be able to rely on their access to domestic courts to enforce what originates as an international legal obligation. Finally, private citizens occasionally have standing to pursue claims before international tribunals, as in the case of the NAFTA investor rights provisions which allow investors access to NAFTA arbitration.

3. Economic aspects of international law

Public international law represents a number of distinct phenomena. Some “law” may be no more than a behavioral regularity in the practices of states, while other law may represent rules coercively imposed on less powerful states by more powerful states.

Still other types of law may arise to promote the domestic objectives of participating officials. But many of the more interesting and important pockets of international law may be seen as efforts to coordinate the behavior of states to address externalities. These externalities may be non-pecuniary, the sort that produce inefficiency in competitive markets, or pecuniary, creating inefficiency due to an absence of competitive conditions.

In this section, I begin with some broad observations about the economic perspective on public international law. Because systematic discussions of international law often imagine that states behave as if they have "preferences," the first topic concerns the conceptualization of states as rational actors. The analysis proceeds to a general discussion of customary international law, to a discussion of the economics of treaties, and finally to consideration of the interface between domestic and national law.

3.1. States as rational actors

Positive economic analysis of international legal regimes conventionally proceeds from an assumption that states behave as if they are rational maximizers over some set of preferences regarding the outcomes of their interaction. The specific assumptions that may be made in this regard are myriad. States may be assumed to behave as economic welfare maximizers, or to maximize a social welfare function that weights the welfare of certain constituencies more heavily than others. The preferences of the "state" may be assumed to be those of its political leaders, who may maximize votes, campaign contributions, or their personal welfare. Innumerable other variations can be imagined depending on the context.

Whatever precise assumption is made about the nature of preferences, it is common to embody a further assumption that states act as if they "care" primarily or exclusively about their own welfare or interests, and less or not at all about the welfare or interests of other states or their political leaders. A divergence will then arise between the *national* maximand and the *global* maximand.

The assumption that states have preference orderings and act as rational maximizers is surely somewhat simplistic. States represent an aggregation of many different actors, whose preferences may well be at odds. The actor with the power to choose among alternatives may change over time, and the constraints imposed on actors with the power to make choices can change over time (in the United States, think of the President as the actor with the power to make choices on international matters, subject to constraints imposed by Congress). Even when it is plausible to assume that a pertinent decision maker has a preference ordering over the available alternatives at a point in time, therefore, the notion that the "preferences" of the "state" are stable over time, or that they obey potentially important regularity assumptions, may be quite problematic.

Although one must acknowledge this problem, there is often little to be done about it in a tractable modeling framework beyond remaining attentive to its possible implications for each subject area. Such a framework proceeds in the tradition of other areas of economic analysis, which embrace their own simple assumptions about the objective functions of corporations, bureaucracies, and other large institutions. Here, as in those other areas, the test is not whether the assumptions are fully descriptive of behavior, but whether they yield useful insights with empirical purchase.

Economic analysis of international law also has its normative side, of course, which rests on assumptions about what states *ought* to be maximizing. Once again, a variety of possible objective functions might be assumed, although the conventional measure of economic welfare is often employed.

3.2. The economics of customary international law and "soft" law

A great deal of work has been done on the economics of "custom" in various contexts. Commentators have written about the use of custom evidence to prove negligence in tort actions, the use of customary business practices as a basis for default rules in contract law, the efficiency of social norms, and the general phenomenon of "order without law" in primitive or frontier societies. Such topics receive significant attention in other chapters in this *Handbook*.

Despite the attention to custom in other contexts, very little has been written about customary international law from an economic perspective. The most notable exception is Goldsmith and Posner (1999, 2005).

Recall the standard characterization of customary international law: it emerges when there is a high degree of convergence in the practice of states, *and* a belief that adherence to the practice has become a legal obligation. The latter requirement is known as *opinio juris* and is central to the existence of customary law according to traditional doctrine. Mere regularities in state behavior, without *opinio juris*, are not law.

Goldsmith and Posner contend that this description of customary international law is largely incoherent. Their alternative theory begins by offering a positive theory of convergence in state practice, which they suggest may result from four distinct phenomena. The first is simple coincidence of interest, whereby all states behave the same way because it is in their unilateral interest regardless of the choices made by other states. They offer "ambassadorial immunity" as a possible example (although this subject is governed by treaty as well as custom in modern times). States may protect the ambassadors of other states, even in times of conflict with them, because the ambassadors perform a valuable function in facilitating communication with other governments. A second explanation for convergence of practice is pure coercion. Here, they suggest that the custom of "free ships, free goods," whereby all property on neutral ships is immune from seizure (including enemy property), is at times illustrative. Powerful states may respect the principle because the seizure of neutral ships to capture enemy property is not worth the bother, while weaker states may respect the principle for fear of retaliation by powerful states. The third possible reason for convergence arises when a common practice represents the solution to an iterated Prisoner's Dilemma, which can be sustained over time by states with open-ended time horizons and sufficiently low discount rates. They again offer ambassadorial immunity as a possible illustration, suggesting that an exchange of ambassadors amounts to an exchange of hostages, and that the prospect of retaliatory acts against a nation's own ambassador abroad can dissuade any temptation to interfere with the ambassadors of others. Finally, convergence may arise in the face of a pure coordination problem where a "focal point" is useful. They suggest that the convergence on a three-mile limit for territorial waters is an example here. For a variety of reasons including security, nations have an interest in claiming dominion over waters along their coast, but the exact limits of territorial waters is to a degree a matter of indifference. A three mile limit supplies a focal point that all nations can accept.

In short, Goldsmith and Posner argue that convergence in state practice occurs for reasons of pure national self interest, albeit not the same reason every time. They further suggest that continued adherence to customary practice happens because the self-interested reasons for convergence remain in place, *not* because of any independent sense of legal obligation. *Opinio juris*, they suggest, is a fiction, and what legal scholars refer to as customary "law" is really no more than a descriptive account of certain regularities in the behavior of states.

To bolster this latter claim, Goldsmith and Posner document how ostensible rules of customary law are frequently violated when states have an interest in deviating. They further illustrate how rogue states, which they suggest have shorter time horizons and higher discount rates, are more likely to deviate than others. Because historical violations and breakdowns of custom can be linked to self-interested reasons for them, Goldsmith and Posner find anecdotal empirical support for the claim that customary practices are mere regularities of self-interest, and that customary law *per se* exerts no tug on state behavior.

The proposition that "customary international law" emerges from the self-interested interaction of states, and that it promotes their mutual interest for one reason or another, seems rather unremarkable. It would indeed be odd if a customary practice emerged on a large scale that made its adherents worse off over an extended period of time. While this aspect of Goldsmith and Posner's analysis seems compelling, a skeptic might argue that they have not fully made their case on the nonexistence of *opinio juris*. Even if customary international law had some force of its own quite apart from the narrow self-interest of a state regarding a particular custom, one might still observe the same anecdotal bits of evidence that Goldsmith and Posner catalog. Nations might still deviate when their self-interested reasons were strong enough, for example, and rogue states might still be the most likely to deviate. All that would be required is that the behavioral force of *opinio juris* be limited, so that counter-incentives of sufficient strength could override it. Thus, although Goldsmith and Posner are surely right that the traditional scholars cannot prove the existence of *opinio juris* by pointing to conformity with custom, neither can the detractors of the traditional view prove its nonexistence merely by pointing to self-interested deviations from custom.

If the empirical evidence is inconclusive, it remains to ask whether *opinio juris* can be given any theoretical content. Why would states feel any obligation to observe a custom that is no longer in their self interest? Traditional international law scholars suggest that once a practice becomes "law," it infuses the morality of national bureaucrats, who then feel a sense of obligation to obey it. One might restate the proposition as a suggestion that "law" has expressive force and alters the preferences of pertinent national actors, leading them to prefer to obey it (or that they simply have an exogenous preference to obey all "law.") The difficulty with this account, as even non-economic scholars have noted, is its circularity. Law exists only after the sense of legal obligation arises according to the definition of customary law, yet the sense of legal obligation is said on this account to follow after the emergence of "law."

A possible alternative account is suggested by Guzman (2002a), who relies on the idea that violations of international law may damage a state's *reputation*. He suggests that international strategic interaction on narrow issues is generally embedded within larger games, and that players' willingness to cooperate with other players on current issues may then turn on whether a player has developed a reputation for cooperation in the past. Under the usual assumptions that prevent backwards unraveling of cooperation (an infinite or open-ended time horizon) and that limit the short terms gains from defection (such as a low discount rate), Guzman argues that reputational considerations create the possibility of an equilibrium in which mutual cooperation is sustained over time in what might otherwise appear to be a one-shot game with defection as the Nash outcome.

The addition of reputation to the analysis suggests a possible economic interpretation of *opinio juris*. One might define it simply as a tendency to obey customary law due to the damage that defection does to a state's reputation as a cooperator, costs that are incurred not in the simple game in which defection is contemplated but in all other games where reputation affects the strategies played by other states.

Traditional international law scholars will find little solace in the reputational interpretation of *opinio juris*, however, because concern for reputation is no less self-interested than concern for payoffs in a narrower strategic interaction. Further, reputational considerations may be of minimal significance as a practical matter in many settings as both Goldsmith and Posner and Guzman argue. This general issue receives further attention below.

Aside from its examination of *opinio juris*, the law and economics literature makes a number of other useful points about the role of customary law. The commentators seem to agree that the ability of customary international law to orchestrate cooperation is limited to narrow circumstances. Problems that require complicated solutions are unlikely to be solved by implicit cooperation—express negotiation and communication will probably be necessary. Further, problems that require the simultaneous cooperation of large numbers of nations will also be difficult to solve because of free rider problems in the enforcement mechanism. Even when a practice appears "customary" on a global scale, therefore, and is thought to represent mutual cooperation, the suggestion is that it is usually no more than a recurring regularity of bilateral interaction.

Guzman makes the further point that if reputation is what creates some "force of law," then there is no reason to limit our conception of "law" to customary international law and treaties. Reputational concerns may be quite important to a world leader who gives her word to another, whether or not it is done in any formal fashion and whether or not it concerns some practice that is widespread in the international community. The traditional line between "hard law" (binding treaties and customary law) on the one hand and “soft law” (such as informal agreements and statements of intention) on the other may thus be quite misleading. Depending on context, states may be considerably more likely to comply with soft law than with hard law, and there is no reason to think that hard law is always preferable for orchestrating cooperation.

In short, economic thinking about customary international law calls into question the very meaning of the concept. It suggests that practices termed “law” in various quarters are no more than behavioral regularities that emerge from self-interested interaction between states facing similar problems. The codification of customary law merely serves to publicize focal points, and to write down the rules of the game to facilitate future adherence to them. The capacity of customary “law” to solve important problems that require cooperation or coordination is quite limited, and will tend to be restricted to issues that admit of a simple solution that can be sustained through small numbers strategic interaction.

Formal modeling bearing on these issues is in its infancy. Fon and Parisi (2003) offer a simple model of custom formation that supports the intuition that customary law is more useful when the preferences of states are relatively more homogeneous. They also consider the role of what they term the “persistent objector’ and “subsequent objector” doctrines that allow states to obtain an exemption from customary rules.

3.3. The economics of treaties and other international agreements

Virtually all of the economic writing on treaties focuses on particular subject areas, with the notable exceptions of Goldsmith and Posner (2005) and Guzman (2002a) cited earlier. In this section I draw to a limited extent on those two sources, but also on ideas developed in more specialized contexts to suggest some general points about the economics of treaties.

In contrast to customary international law, which can emerge through convergence of practice without much communication across states, treaties always involve direct communication, negotiation, and the embodiment of the results in a document. This process is costly, and the reasons for the creation of treaties are narrower or at least different from the reasons given earlier for convergence of state practice on custom. The coincidence of interest explanation for some customary practices, for example, cannot explain why states would incur the costs of creating a treaty. The exercise of pure coercion does not require a treaty either, although to be sure a treaty may be used to orchestrate an end to coercion. Treaties are likely to be valuable instead when state action creates externalities for other states, and when purely decentralized cooperation without formal communication is inadequate to address them (although a few treaties may have other functions, as discussed in later sections).

The mere fact that cooperation is better orchestrated through a process of direct communication, of course, is not sufficient to justify a treaty. Much communication between states occurs without any resulting agreement, and international agreements can arise in the course of communication that are informal and never rise to the level of a treaty. Goldsmith and Posner (2005) thus consider the question of why states resort to "legalization" by formally executing a treaty and making it "binding" as a matter of international law in preference to reliance on less formal, nonbinding agreements. They suggest that the legalization of an agreement may reveal information about a state's commitment to the agreement—in their terms, it shows that the state is "serious" about the agreement. A signal of "seriousness" will only be needed when "seriousness" is private information, and will only be credible if reputational penalties are greater for the violation of a "binding" agreement than for violation of an informal agreement. Hence, this explanation requires that reputation be important to state actors. A second consideration affecting the choice to legalize is the fact that informal agreements may bypass domestic constitutional constraints on the creation of treaties. In the United States, for example, the President has the capacity to conclude and execute informal agreements without Congressional oversight in many areas. Formal treaties (or Executive Agreements that must be approved by Congress) give the legislature greater opportunities to participate and may then constrain the President to less preferred options. But legislative participation may also give the agreement greater durability against changes in administrations, as well as greater force in domestic law. The President will choose between the two options depending on the balance of competing considerations in each case. A final consideration is that legalization subjects the treaty to the interpretive default rules of the Vienna Convention on Treaties. Informal agreements may be chosen out of a desire to opt out of those rules.

Leaving aside for now the choice between formal and informal agreements, it is perhaps useful at this point to set forth a simple, general framework for modeling the potential gains from international cooperation. Variations of this basic approach pervade the literature on individual topics. Imagine two states, denoted A and B, each of which have control over a vector of policy instruments, α and β, respectively. [Nothing changes importantly (beyond the algebra) if the analysis is generalized to N states.] The respective welfare functions for the two states are $W^A(\alpha, \beta)$ and $W^B(\alpha, \beta)$. Assume that each state's welfare is increasing and concave in its own policy choices. The vectors α and β can represent a myriad of policy areas—tariffs, tax rates and rules, immigration restrictions, emissions controls, and so on. In the absence of communication and agreement, each state maximizes its welfare taking the actions of the other as given, selecting α and β such that

$$\partial W^A(\alpha, \beta)/\partial\alpha = 0 \qquad \text{and} \qquad \partial W^B(\alpha, \beta)/\partial\beta = 0.$$

Equilibrium (Nash) arises when both conditions hold, given the other state's choice of policies. Will the equilibrium be (first-best) efficient? The answer is plainly no in general: A point on the Pareto frontier may be derived by choosing α and β simultaneously to maximize the welfare of one state, subject to the constraint that the welfare of the other achieve some fixed, attainable value (a standard technique for deriving conditions for any optimal contract). The first order conditions for this problem require that

$$\begin{aligned} &\partial W^A(\alpha, \beta)/\partial\alpha + \lambda\partial W^B(\alpha, \beta)/\partial\alpha = 0, \qquad \text{and} \\ &\lambda\partial W^B(\alpha, \beta)/\partial\beta + \partial W^A(\alpha, \beta)/\partial\beta = 0, \end{aligned}$$

where λ is a Lagrange multiplier.

With the welfare constraint binding and thus $\lambda > 0$, it is clear that the conditions for Pareto optimality cannot correspond to the earlier conditions for Nash equilibrium *unless*

$$\partial W^B(\alpha, \beta)/\partial\alpha = 0 \qquad \text{and} \qquad \partial W^A(\alpha, \beta)/\partial\beta = 0.$$

In words, the equilibrium without international cooperation will achieve the Pareto frontier only in the absence of externalities, and the function of international agreements in the presence of externalities is to enable states to commit to behavior that will move them closer to the Pareto frontier.

3.3.1. Factors that facilitate or impede international agreement

If externalities suggest an opportunity for international agreements to improve on the equilibrium without them, it does not follow that states will succeed in achieving agreement. Casual empiricism suggests that useful agreements have been reached in a number of areas (e.g., trade), but that many areas laden with apparent externalities have not been successfully addressed through international agreements (e.g., immigration). In other areas, some issues have been addressed through agreement but not others (e.g., investment and environment). What explains these various "successes" and "failures"?

Perhaps the starting point for analysis is the Coase Theorem. One would expect international agreements to exhaust potential joint gains from solving externality problems only to the degree that those gains remain after all transaction costs have been accounted for in the calculus (which must be understood broadly to include not only monetary costs of achieving agreement but all factors that affect the political acceptability of agreement). To understand the universe of international agreements (and the areas in which they are absent), one must therefore ask not simply whether externalities arise, but also whether the transaction costs of international agreements to address them are low enough to permit agreements to go forward. A variety of considerations will affect the magnitude of transaction costs.

Trivially, agreement can only arise if some agreement lies in the core of the bargaining game—each state that becomes party to an agreement must perceive itself better off than by refusing to participate (or by breaking off with others into a smaller numbers agreement in the multilateral case). It is easy to imagine settings in which international externalities lead to an equilibrium off the Pareto frontier in the absence of cooperation, yet where the bargaining options are too limited to admit of a Pareto improvement in the core. For example, imagine two states, one of which contains a monopoly producer and exporter of widgets, and the other of which is a consumer of widgets with no monopoly power over any tradable good or service. The monopolist exploits its market power with the familiar deadweight losses, although much of its monopoly profit comes as a transfer from consumers abroad. Global welfare would increase if each state pursued a sensible anti-monopoly policy. But if the two states try to strike a bilateral agreement that merely requires each of them to adopt an anti-monopoly policy, the state with the export monopolist may well object that such an agreement leaves it worse off than without it.

Two obvious and related solutions to the problem suggest themselves. The first is monetary side payments from one state to the other. This device is sometimes employed in practice, usually in the form of a promise of monetary aid to a state that cooperates on some issue (the recent aid to Pakistan associated with its quiet assistance in the invasion of Afghanistan is illustrative). Second, and probably more common in practice, the scope of bargaining can be expanded to include other issues (or perhaps other states, a point explored later). The state that enjoys a widget monopoly may be willing to forego monopoly rents in exchange for valued concessions on security issues, environmental matters, and so on. The general lesson is that issue linkage in international negotiations can greatly expand the scope of possible agreements. A likely recent example is the Agreement on Trade Related Aspects of Intellectual Property (TRIPs) in the WTO. Developing nations, which are primarily consumers rather than producers of intellectual property, were induced to agree to strengthen their intellectual property laws in ways that would confer considerable rents on foreign rights holders in exchange for concessions on other trade issues such as textiles and agriculture.[2]

Issue linkage in particular has received some formal attention in the literature.[3] Horstmann, Markusen, and Robles (2001) consider a two-party Nash bargaining game involving the division of surplus associated with two "issues" of concern to the bargainers. One bargainer derives greater marginal utility from its share of the surplus on the first issue, while the second derives greater marginal utility from its share of the surplus on the second issue, creating a difference in their "comparative interest" in concluding a deal on each issue. Horstmann, Markusen and Robles show that in an "unlinked" bargaining arrangement (two simultaneous games in which an offer on one issue cannot be conditioned on the outcome of bargaining on the other or made dependent on the other), the parties will fail to achieve their utility possibilities frontier. The reason is that the Nash bargaining solutions for the separate games will fail to take advantage of each party's preference for greater surplus from one game over the other. When the bargainers can "link" their offers by making a simultaneous offer over both sources of surplus, by contrast, the utility frontier is achievable because each party can reap a greater share of its preferred surplus (each party receives more on the issue of greater comparative interest).

It is too optimistic, however, to imagine that either monetary side payments or issue linkage will always eliminate the problem of an empty core. Take the monopoly hypothetical above and consider the monetary side payment option. If an anti-monopoly

[2] Although I use TRIPs as an illustration of issue linkage here, I do not mean to imply that TRIPs necessarily enhanced global welfare conventionally defined. Here, as in many other settings, one must distinguish carefully between political optima that maximize the interests of parties to negotiations, and conventional welfare optima.

[3] On the normative economics of issue linkage in the trade area, with particular reference to competition policy, labor and environmental standards, see Bhagwati and Hudec (1996).

agreement is to lie in the core, the consuming state must make a monetary payment to the state with the monopolist that is large enough to induce it to give up its monopoly rents. Such a payment may leave little gain to the consuming nation in relation to the transaction costs of negotiation. As for issue linkage, an expansion in the scope of negotiations inevitably increases their cost, and draws in a greater number of domestic political constituencies with an interest in the outcome. The resulting increase in the costs of international negotiation and of domestic political deliberation may be great enough to undermine the process. Nevertheless, many international agreements display a great deal of issue linkage in their formation, as prominently illustrated by the WTO and NAFTA.

Moving beyond the problem of ensuring that an agreement lies in the core, the transaction costs of reaching an international agreement turn importantly on the *complexity* of the issues to be addressed. Some agreements entail reasonably simple commitments (such as the ban on nuclear testing in the Nonproliferation Treaties). Others require highly complex commitments spanning a wide array of issues. The WTO is again illustrative. To make its central commitments on tariff reductions valuable, signatories must disable themselves from turning to substitute instruments of protection. The result is hundreds of pages of treaty text addressing quotas, import licensing restrictions, balance of payments policy, domestic tax and regulatory policy, state trading and monopolies, and many other subjects. Complexity can be said to increase not only as the detail involved in specifying the proper behavior of a state increases, but also as the optimal behavior across states becomes more variable. Both dimensions of complexity raise the costs of agreement not only because desired behavior becomes more costly to describe and memorialize, but because enforcement may become more difficult as the number and variety of possible violations multiplies.

A third important factor affecting the transaction costs of reaching and enforcing an international agreement is the number of countries involved in it. Even holding constant the complexity of the commitments (which of course may increase as more states participate), greater numbers of states add to the negotiation costs of reaching agreement. Logistical costs increase because of the larger numbers, and delays may develop as states hold back on what they offer to achieve agreement hoping to free ride on inducements offered by other states. A free rider problem may also manifest itself in the enforcement mechanism. Imagine that breach of agreement requires some action by non-breaching parties to punish the party in breach. If these actions are costly, each state may prefer that others do the punishing, and unless some coordination mechanism can be designed to overcome the problem the threat of punishment may lose credibility. This observation suggests an important distinction between situations in which punishment requires actions that the punishers view as costly to themselves, and situations in which the opportunity to punish another state is welcomed. An example of the first may be military force. A state that employs military force risks the lives of its troops and consumes monetary resources, and most states will be happy to have others undertake the task if they can perform it as well. By contrast, imagine an environmental agreement limiting emissions of some pollutant in each state, enforced by a threat of mutual defection. Here, should another state defect, a punishing state might well benefit economically from the opportunity to defect itself and there may be no free rider problem in enforcement.

These last observations suggest some further principles. If externality problems can be handled adequately on a bilateral basis, such arrangements will tend to be preferred to multilateral arrangements. The case for agreements involving larger numbers of states maps directly onto the case for international agreements in the first instance—multilateral cooperation is potentially desirable mainly when bilateral cooperation itself creates externalities for third states. The caveat is that if a multitude of bilateral agreements would all look about the same, it may be more economical to create one multilateral agreement that sets out the same principles in a single document.

A related point is that when bilateral or small numbers agreements create externalities, an expansion of the number of states participating in an agreement will not simply increase the costs of agreement, but also the benefits. Hence, the international community may move toward large scale multilateral agreements despite their higher cost.

Finally, not only does the number of states involved affect the costs of achieving and enforcing agreement, but the size distribution of states may also matter. In particular, the presence of a few large states may help at least on the enforcement problem. If punishment is seen as costly to the punishers, the free rider problem will be less acute when some states are large enough to capture a considerable portion of the joint gains from enforcing the agreement. It may then be in their private interest to act as enforcers even if smaller states will free ride. This possibility suggests a constructive role in some contexts for what political scientists term a "hegemon," a powerful state that enforces its will in an environment populated by other smaller and less powerful states.

3.3.2. Treaties as contracts

As noted, one may model many international agreements as an effort by states to reach their Pareto frontier in the presence of externalities. They do so subject to a participation constraint for each state and perhaps other constraints depending on the context.

Readers familiar with the literature on optimal contracting will find this type of problem quite familiar. Indeed, from an economic standpoint, a treaty *is* a contract, albeit one between states rather than individuals. As long as one is prepared to reduce the "preferences" of the state to an ordering akin to that of an individual rational actor, there is no difference in general between the problem of designing an optimal contract and the problem of designing an optimal treaty. Many of the ideas in the optimal contracting literature thus have direct bearing on the study of treaties. A few illustrations follow.

Treaties as incomplete contracts Some treaties address simple matters about which optimal behavior changes little over time. But many treaties address complex matters in an environment subject to uncertainty. For familiar reasons, treaties in the latter group are likely to be incomplete as to certain behaviors and contingencies. States may then benefit from various devices to fill the gaps in their agreement. Specialized tribunals may supply useful default rules in some cases, as may overarching treaties on interpretation such as the Vienna Convention on Treaties. As in the literature on default rules for private contracts, one can ask what the guiding principle should be in designing them. Should they replicate what the parties would have negotiated expressly if they had addressed the matter, penalize a party that has withheld pertinent information, or something else?

The contractibility of desired behavior It is a commonplace in the literature on optimal contracting to suppose that some behavior is non-contractible, usually because the behavior in question cannot be observed by other parties, or at least cannot be verified by a court. The same class of problems arises under many international agreements regarding everything from hidden non-tariff trade barriers to cheating on nonproliferation commitments. Treaties may then employ the familiar types of (generally second-best) solutions. Treaty obligations may be conditioned on verifiable behavior that is correlated with the unverifiable behavior, as in the classic principal-agent models where payoffs must be conditioned on observable outcomes rather than unobservable effort. Alternatively, provisions may be designed to encourage decision makers to internalize the externalities from their choices, and thereby to eliminate the divergence between privately and socially optimal behavior.

Renegotiation, modification and efficient breach Treaties are often negotiated under conditions of uncertainty. A variety of shocks may cause particular commitments to become inefficient, or may leave some signatory worse off than it would be by exiting. Some treaties will address the problem simply by providing for the possibility of withdrawal, perhaps after a period of notice (the SALT treaty, for example). But many treaties address a broad range of issues, and changed circumstances may justify only a modification of the bargain, not a complete end to it. Accordingly, treaties may contain provisions providing for the renegotiation of parts of the bargain or may specify contingencies under which states may deviate from their prior commitments. A treaty may provide that a party can breach its obligations at a price, which if set correctly can facilitate "efficient breach."

The design of such provisions can raise a number of interesting questions. In a multilateral treaty, for example, how does a renegotiation avoid the holdout problem? Does that problem argue for allowing breach at a "compensatory price?" How is the price for breach set, and when is it preferable to employ stiffer sanctions that move the regime toward a "property rule" and away from a "liability rule?" These latter questions presuppose that the legal regime has the capacity to employ meaningful sanctions for breach and to calibrate them—a subject that will receive further attention in a moment.

In addition to the problem of facilitating efficient adjustments to changed circumstances, treaties must confront the fact that efforts to modify the bargain may be opportunistic. In the event that a party deviates from its commitments and then offers to renegotiate, will it be in the interest of the other parties at that point in time to hold it to the original bargain and will they have the ability to do so, or will they capitulate to new terms that extract some of their surplus? In many important contexts, therefore, a treaty designer must worry whether the agreement is renegotiation proof. This problem is usually related to the presence of sunk costs. After states have made sunk investments in reliance on an agreement, the returns to those investments may be vulnerable to expropriation during opportunistic renegotiation.

This is but a partial listing of the issues that international agreements must confront, and that have direct parallels in optimal contracting problems. Economically-oriented scholars will find many fruitful applications to international law of the ideas developed in the contract literature.

3.3.3. Enforcement and dispute resolution

States contemplating an international agreement often confront another set of problems that is absent for parties to conventional contracts. Private actors commonly rely on state enforcers to hold them to their commitments. An award of damages in contract can be enforced through a seizure of the promisor's assets or an injunction backed by threat of imprisonment. By contrast, although the use of military force and the seizure of assets is surely seen in international relations, such devices are not part of the enforcement arsenal for many international agreements. Indeed, the observation that much of international law is not backed by a credible threat of military force or other strong coercive measures has led many commentators to question whether international law is "law" at all.

Such skepticism is misplaced in my view for at least two reasons. First, states that join international agreements can and do create international regimes where strong coercive measures are possible—witness the occasional deployment of troops after U.N. authorization through the years, or the various episodes of U.N. sanctions that have affected the economic vitality of rogue states. International law should not be viewed as hopelessly weak because of a limited enforcement regime that is imposed on it exogenously. Instead, one must recognize that the choice of enforcement regime is endogenous. There is much that states can do, if they wish, to make their commitments more credible. One must therefore ask why some international legal commitments carry much greater punishments for breach than others, and whether weaker regimes of enforcement are inadequate to their task or are instead chosen precisely because nothing more is necessary.

Second, although parties to private contracts can often appeal to state enforcers, in many settings they cannot. Litigation costs may swamp the benefits of an appeal to the courts when the monetary stakes are modest. Transnational contracts can confront difficult issues of securing personal jurisdiction in the event of a dispute, enforcing awards, and securing unbiased decision makers. Such problems have led economic scholars to recognize that there are valuable mechanisms for making contractual commitments credible that do not rely on the existence of a third-party enforcer with coercive powers. (See, e.g., Telser, 1980). Even when these mechanisms cannot achieve the "first-best," they can often accomplish a great deal, and so too at the international level.

I thus offer a brief catalog of the mechanisms that are available for the enforcement of international agreements that do not rely on the coercive powers of third party enforcers. Some have been mentioned in passing already but deserve more detailed comment.

Hostage exchange and bond posting The paradigm example of the hostage exchange involves a hypothetical peace treaty between warring kings. Each king sends a son to live in the other kingdom, with the understanding that if he attacks that kingdom his son will be killed. Both kings value their sons' lives more than any possible gains from aggression, and so peace is sustained.

The exchange of human hostages threatened with death is rather unseemly by modern standards, of course, and we do not observe it in international agreements. But analogous situations may arise, as when foreign nationals or their assets travel and become subject to the jurisdiction of other states. When such movements are reciprocal a situation akin to a hostage exchange may arise implicitly—recall the suggestion by Goldsmith and Posner that the rules of ambassadorial immunity may be sustained through such a mechanism. Another rough analogy arises in the trade area, where trade liberalization may lead to mutual specialization (and sunk costs) in industries that are dependent for their viability on continued access to the market of a trading partner. See Devereux (1997). A possible difficulty with the hostage exchange mechanism, of course, is that the reciprocal threats to the hostages may not be credible. It may not be in the interest of the king who is attacked to kill the attacker's son, for example, if he knows that his own son will then be killed.

A related mechanism is bond posting. Each party posts a bond that is large enough to exceed its potential gains from cheating on the agreement. Should cheating occur, the bond is forfeit. Compliance with the agreement then becomes an equilibrium as long as cheating will be detected with sufficient probability in relation to the size of the bond. Bond posting differs from hostage exchange in that it generally relies on a third-party arbiter to determine whether breach of agreement exists, but the arbiter need not possess any coercive powers beyond the capacity to declare that one party has forfeited its bond to another. If the arbiter is trusted by all parties, and they cannot interfere with the exercise of the arbiter's authority, then a bond posting mechanism can do quite well in encouraging compliance. Interestingly, formal bond posting arrangements seem quite rare in international law, although the reason why is by no means obvious.

Mutual threats of defection International agreements that address externalities often have many of the qualities of a repeated Prisoner's Dilemma. Each party makes commitments that it would prefer not to make, other things being equal, in exchange for valuable commitments from others that result in a Pareto improvement. States will be tempted to defect if they think they can do so without retaliation.

Repeated Prisoner's Dilemmas have been studied extensively, both theoretically and empirically. On the theoretical side, the well known Folk Theorem holds that long term cooperation is a possible equilibrium, enforced by mutual threats to defect from cooperation should another party defect. Cooperation requires that the game have no fixed ending (defection would become a dominant strategy in the last period and cooperation would unravel from there), and that the parties have low enough discount rates that the current gains from defection do not loom too large in relation to the long term gains from cooperation. Rasmusen (1989). A requirement that the equilibrium be renegotiation proof makes the equilibrium rather complicated, but such equilibria do exist (Fudenberg and Tirole, 1991). Empirical research suggests that the "tit for tat" strategy, whereby a period of defection by one party is met by a subsequent period of defection by others, can be a successful enforcement device, even if the theoretical literature notes that this strategy is not subgame perfect. Axelrod (1984). Although much of the discussion in the literature is of 2-person repeated Prisoner's Dilemmas, many international agreements represent the N-person variant. Here too, cooperation is a possible equilibrium but hardly the only one.

Regardless of the number of players, threats of mutual defection are more likely to sustain cooperation when defection by one party can be detected readily by others. It is also important that parties be able to agree on what constitutes defection, and hence that the rules of the game be clear. Otherwise, what one party claims to be a justifiable punishment for defection by another party may itself be viewed by others as opportunistic defection.

One important factor that aids cooperation in international agreements is absent from classic models of the Prisoner's Dilemma. Under the usual assumptions about the way the game is played, the parties cannot communicate with each other during a "period" of play, and can only discover what the other party has done after each has made a simultaneous move. Parties to international agreements, by contrast, can remain in close communication with each other at all times. Depending on the context, any movement toward defection may be easy to detect. Indeed, any plans for defection may be public knowledge and even the subject of public debate long before defection actually occurs. Consequently, defection may become punishable more quickly. From an analytic standpoint, improved communication shortens the "periods" during which the repeated game is played, and has a tendency to reduce the short term gains from defection.

International sanctions (unilateral or multilateral) Any punishment for breach of an agreement may be termed a sanction, but here I use the term more narrowly. Define "sanction" as a *costly* measure, other than retaliatory defection, taken by a state aggrieved by breach of an agreement. For example, if state A violates its obligations under a Nonproliferation Treaty, state B might impose a sanction in the form of a suspension of trade relations. Sanctions may be undertaken unilaterally, or in coordinated fashion by a number of states.

So defined, a key feature of sanctions is that they impose costs on the states that employ them. The strategic setting thus differs importantly from that of a repeated Prisoner's Dilemma, in which retaliatory defection can benefit a state after another has already defected (at least in the 2-person case). The question whether to carry out a sanction under these conditions is sometimes termed the "Punisher's Dilemma."

Sanctions too have received considerable study, both theoretical and empirical. Eaton and Engers (1992) model sanctions in a two-country framework with one country as the "target" of sanctions and one as the "sender." The sender makes a demand on the target, which can then balk or comply. If the target balks, the sender can impose a sanction that is costly both to itself and to the target. They show that subgame perfect equilibria exist in which the threat to impose the sanction succeeds and the target complies as long as discount rates are not too high. The sender's threat to impose the costly sanction is credible, they argue, because a reputation for toughness is valuable to it in the future. The equilibria that sustain compliance are not renegotiation proof, however, as following an incident of balking it would be in the two countries interest to forget about past transgressions, so that in the simple case the only renegotiation proof equilibrium is unpunished balking.

They then consider a more complex case in which compliance is a matter of degree. Renegotiation proof equilibria then do emerge in which the threat of sanction induces some level of compliance. The degree of compliance will be greater, other things being equal, the less costly the sanction is to the sender. A high degree of compliance can be exacted if, for example, the sender can extract reparations from the target to cover the cost of the sanction. Further complications arise when there are more than one "sender" country. Depending on the distribution of costs and benefits from acting, some countries may free ride on the sanctions efforts of others, and it is possible that no country will take action.

Chang (1995) considers a somewhat different set of issues—the choice between sanctions and bribes, or "sticks" and "carrots." He develops his analysis with particular reference to global environmental issues, and to the ongoing debate over whether developing nations should be compensated for measures that create positive externalities or that avert negative externalities (such as preserving rain forest or protecting sea turtles). The analysis has broader applicability, however, and bears on any situation where nations are choosing between bribes and sanctions to induce cooperation. Chang's essential concern is that in any regime where "carrots" are offered, nations may react strategically to extract larger "carrots." Drawing from the game-theoretic literature on predatory pricing, Chang develops a signaling model in which states that offer bribes are imperfectly informed about the potential recipients' private optima in the absence of a bribe. There are two "types" of potential recipients, those whose private optima in the absence of a bribe are relatively close to the desired cooperative behavior (good types), and those whose private optima in the absence of a bribe are relatively far from the desired cooperative behavior (bad types). They all signal their type by engaging in some level of the behavior in question (cutting down rain forest, for example). A possible outcome is a pooling equilibrium in which the good types mimic the bad types to extract larger bribes. Such a development is potentially unfortunate, of course, because it raises the costs of securing cooperation and may lead to perverse signaling behavior that exacerbates the underlying problem.

Turning from theory to empirics, the historical efficacy of sanctions has also received considerable study. Any such study must come with an important caveat—we cannot observe sanctions that were never imposed because the mere threat of them achieved desired compliance all along, and we cannot observe cases in which sanctions were not imposed because they were perceived to be futile. Cases where they are actually employed, therefore, must represent intermediate instances where states were uncertain about their efficacy or were simply building a reputation for toughness. A notable study is that of Hufbauer, Schott, and Elliott (1990), who reviewed the use of sanctions in 116 cases since World War I. They found that sanctions succeeded about one-third of the time. They were more likely to succeed, *inter alia*, when the policy change sought by sanctions was relatively modest, when the sanctions were more costly to the target, and when the sanctions were less costly to the sender. A review of theoretical and empirical literature on sanctions may be found in Eaton and Sykes (1998).

Reputation Analytic accounts of reputation rely on the notion that actors have different propensities to cooperate with others, and that their "type" in this respect is private information. Actors prefer to enter cooperative arrangements with other actors who will honor them. An instance of defection from cooperation by an actor thus conveys useful information to other actors about their likely type; those who defect will be judged less reliable in the future, and will lose opportunities for cooperation. When future gains from cooperation are important and discount rates are low enough, the desire to maintain a reputation for cooperating can lead actors to do so when they would otherwise be tempted to defect. Baird, Gertner, and Picker (1994).

The role of reputation in policing international agreements is controversial. Most commentators acknowledge that it likely plays some role, although some imagine it to have a central role while others doubt that it amounts to much most of the time. In general, it is unclear to what extent reputation crosses over from one set of issues (say, trade) to another (say, military security). Even within an issue area, the notion that the propensity of states to cooperate is "private information" seems questionable, particularly in the more transparent Western societies. It is also unclear how well reputation survives through time, especially as political leaders change, raising the possibility of "end game" issues. The importance of reputation is thus unclear.

Information revelation mechanisms All of the enforcement devices noted above rely on the ability of states to detect violations of international agreements. Sometimes violations are obvious, as when a nuclear test occurs within the territory of a signatory to a Nonproliferation Treaty. But sometimes violations are surreptitious, and in other cases disagreement may arise over what constitutes a violation. In response, a number of devices may be employed to improve on the information possessed by parties to an international agreement.

If violations are surreptitious, a need may arise for investigators to examine suspected violations or to conduct routine inspections for compliance. Parties may place greater confidence in neutral investigators who are not linked to disputants, and a role for an institution with independent investigators may emerge. U.N. weapons inspectors are perhaps a useful illustration of this possibility.

Even in cases where pertinent behavior is observable by other states without investigators, a mechanism may be needed to resolve factual and legal disputes regarding alleged violations. None of the mechanisms outlined above can work well unless states know whether observed conduct is compliant or deviant. Here, a role emerges for a formal tribunal to conduct factual and legal analysis. Related, it may be valuable for violations to be publicized extensively. Tribunals and their decisions, if public, can also serve this role. Requirements for centralized reporting of violations, even in the absence of a tribunal, can serve a valuable publicity function.

3.3.4. Voting

Some international agreements and many international organizations require collective decision making on various issues. A question arises as to how such decisions shall be made. Some are taken during a process of renegotiation, as has been seen on many issues throughout the history of the WTO/GATT system. Others are taken by vote. If a voting mechanism is to be used, the parties must select a voting rule.

Many voting rules might be employed, each with obvious advantages and disadvantages. A unanimity rule, which is not uncommon in international organizations, ensures that no state must accept policies with which it disagrees. But it also raises a substantial impediment to change, which may reduce group welfare. Rules such as majority voting have roughly the opposite problem. They may be more helpful in facilitating change that is valuable to the members of the organization as a whole, but they run the risk of requiring substantial numbers of participants to accept policy changes that injure them, and thereby raise the prospect of the organization unraveling through defections.

Maggi and Morelli (2003) address this set of issues formally, on the assumption that collective decisions must be self-enforcing. They show that in circumstances where defections are of greater concern, as in organizations where the members behave as though their discount rate is low, unanimity rules will be more attractive. Majority rules tend to be more attractive in the opposite set of circumstances.

3.4. The interface between domestic and international law

A number of interesting issues arise regarding the relationship between domestic law and international law. Here, I will focus on two areas: the problem of making negotiators credible in the face of domestic political constraints on their power, and the process of conforming domestic law to international obligations.

Putting aside customary law, international law emerges only after a process of state to state negotiation. To reduce transaction costs, it is useful for each state to speak with "one voice" in the process. But it is not uncommon for power over the issues in question to be shared among many domestic actors. The United States is a good example, and I will use the trade area for concreteness. Although the President and his trade officials are the natural parties to negotiate on behalf of the United States, the constitutional separation of powers requires that virtually all changes in trade policy be approved by both Houses of Congress. Indeed, all tariff bills must originate in the House Ways and Means Committee. As a result, certain key legislators and committees have a great deal of power over trade policy, enough to block or bottle up trade agreements negotiated by the President if they wish.

Because of this problem, it is conventional wisdom that the President should not embark on trade negotiations without benefit of the so-called "fast track" authority. The fact track amounts to a procedural rule, adopted by both Houses, promising that any proposed trade agreement submitted to Congress will be voted on within a fixed timetable (powerful committees cannot delay the vote), and will be approved or disapproved in its entirety with no opportunity to offer amendments. The last aspect of the rule ensures that Congress cannot condition its approval on some item that would require renegotiation.

The economics of the fast track are interesting. A possible explanation is that legislators expect trade deals to benefit themselves ex ante, but also know that ex post some powerful figures will likely wish to block approval. Because of this prospect, they know that agreements are unlikely to be approved without the fast track, and indeed other countries may not bother to negotiate. They agree on fast track "behind a veil of ignorance," uncertain as to who will end up regretting it ex post but secure in the knowledge that the expected benefits to those who vote for it are positive ex ante.

Although plausible, this explanation is not without difficulties. It relies heavily on the notion that uncertainty prevails ex ante as to which constituencies will suffer under a trade agreement. Yet, the negotiating agenda of other trading nations is usually quite well known before negotiations begin, and the industries likely to be imperiled by further trade liberalization are not terribly difficult to identify. The mystery then becomes why the fast track procedure does not simply cause the political bottleneck to emerge at the time of the fast track vote, when the powerful figures who would stand in the way of a final agreement instead make their stand against fast track authority. A possible retort is that the political costs to legislators who refrain from blocking fast track authority are somehow lower than the costs that they would face should they retain the power to block an agreement ex post and fail to exercise it. This explanation requires constituents to be ill-informed to a degree, and unable to discern that their plight ex post was highly predictable ex ante. Hence, it is also a somewhat uncomfortable explanation. Leebron (1997) offers a thorough discussion of fast track, accompanied by skepticism that it really makes much difference.

The fast track example is but one illustration of a broader class of problems that arises with respect to many types of international agreements—how to make the actors who negotiate on behalf of each nation credible so that their representations may be relied on, while ensuring that they act as faithful agents for their principals. This class of problems has received very little study by economically oriented scholars.

Another interesting aspect of the interface between domestic and international law concerns the status of international law in domestic legal systems. With respect to treaties, it is understood that they may be "self-executing" or "non-self-executing." A self executing treaty is intended by the signatories to be incorporated directly into their respective legal systems. As a result, the treaty is designed to become immediately enforceable in domestic courts—it may thereby create private rights of action, or trump prior inconsistent domestic law. A non-self-executing treaty is a binding international obligation but has no domestic legal effect *per se*. A party to such a treaty can choose to give it effect in domestic law or not through further domestic legislation or legal action (of course running the risk that it may violate international law if its domestic law is not properly conformed).

A closely related distinction exists between "monist" and "dualist" states. A monist state takes the position that all binding international legal obligations are incorporated into domestic law automatically. In effect, international law is always "self executing," even when other states do not treat it as such. A dualist state takes the position that international law has no direct domestic effect, except in particular cases where a mutual agreement exists that it should be self-executing. Absent such a case, the dualist state retains the right to implement the international obligation into its domestic legal system as it sees fit, again subject to the risks and consequences of violating international law. The United States, for example, is a dualist state. Except for rare self-executing treaties, international law becomes effective domestically only through additional (usually Federal) legislative enactments. By contrast, Costa Rica is a monist state, and even goes so far as to elevate international law above its own constitution.

At first blush, dualism seems to create a great danger of opportunism. Many international agreements require extensive changes in domestic law before states can bring their behavior into compliance. The dualist state may promise to make the necessary changes, but the task of monitoring compliance at all levels of government may be a daunting one for other states. The dualist state may then fail to behave as promised without detection for a long time, and one might think it better for international law to have direct domestic effect to eliminate the problem. This view of dualism is too harsh, however, for it neglects the fact that conforming domestic laws and practices is costly, whether accomplished through domestic legislation or through the application of new international law. And on many issues, strict compliance with international law may be of little importance to other states—U.S. trading partners may care little about whether Alaska conforms its building codes to the WTO Technical Barriers Agreement, for example, if the affected trading volume is nil or de minimis. Dualism allows states to incur the costs of converting their domestic laws only as they have reason to believe that other states care about the conversion, and may thus afford Pareto gains to the participating states. Whether these gains offset the problem of opportunism in any given setting is an empirical question, and where the opportunism problem is obviously dominant the dualist states can avoid it case-by-case through the occasional self-executing treaty.

The harder question concerns the existence of some monist states. Why would any state subject itself to the direct effect of international law unless other parties to an agreement promise to do the same? The answer here may lie with transaction costs. The costs of enacting domestic legislation to implement international obligations may be partially fixed, so that such costs represent a larger proportional tax on a small economy than on a large economy. At some point, the costs of domestic lawmaking may exceed the benefits of the selective deviation from commitments that dualism permits. The smaller state may then prefer to embrace the international law wholesale rather than to incur the costs of determining how it can benefit from a dualist approach.

It is also conceivable that monism is valuable as a device for making a commitment to domestic firms. A monist state will have greater difficulty changing the law in the future than will a dualist state. When it is in the interest of the government to induce firms to rely on the current state of the law, therefore, monism has its advantages. Of course, a state could perhaps do better by adhering to a posture of dualism in general, and then selectively entering self-executing agreements in the instances where commitment to the law is especially important. On the general phenomenon of monism and dualism, and the variety of intermediate approaches that emerge in various nations, see the collection of papers in Jackson and Sykes (1997).

4. Security issues in international law

This section addresses the international law of military conflict and strategic alliances. These subjects have been the subject of intensive study by political scientists, but relatively little work has been done by law and economics scholars until recently. The exception concerns work on collective defense as an application of the theory of public goods.

4.1. International alliances

Olson and Zeckhauser (1966) offer an early formal treatment of "alliances" that supply defense services to their members as a public good. Various scholars have extended their work, primarily by treating defense as a mixed public/private good. Sandler and Cauley (1975). A recent survey of the literature both theoretical and empirical may be found in Sandler and Hartley (2001).

The essential issues may be captured in a two country model, which generalizes to N countries quite easily. Imagine two allied countries, one and two, each of which must decide how much of its resources to allocate to defense measures and how much to allocate to a numeraire good. Denote units of defense measures in each country by d_i. For simplicity, let the price of defense measures be p in both countries. Each country's consumption of the numeraire good is y_i, and the total income is I_i. The national budget constraint is thus $y_i + pd_i = I_i$.

Defense measures by each country may benefit the other—a nuclear deterrent, for example, may discourage attacks on both nations. But defense measures may also have a rival component to them, as with measures to defend one's own border against terrorist intruders. In this sense, defense measures yield a joint product. Let the rival or "private" component of defense be denoted $x(d_i)$, and the non-rival or "public" component be denoted $z(d_1 + d_2)$. The welfare functions of country one and country two are $U(\cdot)$ and $V(\cdot)$, respectively. Each welfare function has four arguments: consumption of the numeraire good y, the private component of defense $x(d_i)$, the public component of defense $z(d_1+d_2)$, and a shift parameter Ω_i, which may be interpreted as the magnitude of the perceived external threat.

Consider first the choices that each country will make, taking the choice of the other as given. It will suffice to examine country one's problem, as country two's problem is completely analogous. Country one will choose y_1 and d_1 to maximize $U[y_1, x(d_1), z(d_1+d_2), \Omega_1]$, subject to its budget constraint. The first order conditions imply:

$$x'U_x/U_y + z'U_z/U_y = p.$$

The first term on the left hand side is the marginal rate of substitution between the private component of defense and the numeraire good, multiplied by the marginal effect of greater defense measures on the private component of defense. It can be interpreted as the marginal money value of the additional private component of defense generated by an increase in country one's defense measures. The second term may similarly be interpreted as the marginal money value *to country one* of the additional public component of defense generated by an increase in its defense measures. Country one will equate the sum of these values to the price of a unit of defense measures. Country two's optimum is identical, restated using derivatives of its own welfare function $V(\cdot)$. Nash equilibrium occurs when each country's choice of defense measures is optimal given the choice of the other.

The Nash equilibrium is not socially optimal, of course, because each country neglects the positive externality generated by its defense measures through the public component of defense. A social optimum may be found by choosing the y_i and d_i to maximize the welfare of country one, subject to the constraint that the welfare of country two achieve a fixed value, and to the joint budget constraint $y_1+y_2+p(d_1+d_2) = I_1+I_2$. The first order condition for a socially optimal choice of d_1 may be written:

$$x'U_x/U_y + z'(U_z/U_y + V_z/V_y) = p.$$

This expression includes country two's marginal rate of substitution between the public element of security and the numeraire good, and thus captures its willingness to pay at the margin for additional defense measures by country one. The socially optimal choice of d_2 must satisfy an analogous condition. The appearance of the sum of the marginal rates of substitution across countries in the conditions for social optimality is a standard type of result in the theory of public goods. Indeed, if one were to eliminate from the welfare functions the arguments $x(\cdot)$ and thereby eliminate the private component of defense, the model would be a trivial variant of a standard pure public goods model. Conversely, if one were to eliminate the $z(\cdot)$ arguments, the externality would vanish and Nash behavior would be socially optimal.

This type of model may be used to generate a number of implications. For example, if we consider the pure public good case for simplicity and make the conventional assumption that marginal welfare is diminishing in each argument, smaller defense expenditures by one country will induce larger expenditures by the other in Nash equilibrium and vice-versa. An increase in the threat parameter to country one, therefore, which increases its marginal welfare from defense, will lead it to increase its defense measures and lead country two to reduce them (the matter becomes more complicated when security has both a private and public element). Defense measures are also a function of income, of course, and changes in the relative income levels of the two countries will have similar implications. With sufficiently large income disparities, the countries' reaction functions may not cross at all and the wealthier country may shoulder all of the joint defense burden while the poorer country free rides completely. Models of this sort have been offered to explain a number of empirical observations, such as why so much of the cost of nuclear deterrence was borne by the United States during the Cold War. The other key implications arise when we allow the private component of defense to become more important. The greater the extent to which defense measures yield rival instead of non-rival benefits, the lesser the tendency of allies to undersupply defense.

The reader will no doubt notice that while this model comes directly from the literature on "alliances," it is in fact a model of equilibrium behavior in the absence of an alliance (aside from the fact the two countries are "allied" against some of the same threats). The role for a formal alliance in this context is to overcome the collective action problem that arises when defense measures have important positive externalities. Thus, having identified the externality problem, one must move on to ask whether defense policy in practice is importantly affected by it. If so, one must next ask what mechanisms can best abate it, and whether existing alliances have addressed it satisfactorily.

These are difficult questions. As to the first, it is hardly enough to observe that some types of defense measures generate positive externalities for allies because potentially offsetting factors exist. At least since the concept of the "military-industrial complex" was introduced by Eisenhower, a public choice perspective on defense questions whether interest group politics may bias defense expenditures upward. It is thus difficult to say whether the level of defense spending is too high or too low in many contexts, either on an individual basis or across some specific alliance. The judgments of individual commentators about the adequacy of defense often depend on varying and highly uncertain estimates of the magnitude of external threats.

Assuming that members of an existing or potential alliance can agree that their collective spending is too low, however, the opportunity for them to negotiate over their respective contributions may bear fruit. Each nation can offer to contribute more in return for greater contributions by others. Although the incentive to free ride and hold out during negotiations will no doubt present itself, it is possible that mutual commitments to increase contributions will arise and be respected. A useful illustration may be drawn from the history of funding for U.N. peacekeeping. In the early days, the organization would solicit voluntary contributions, which were predictably small and few in number. As peacekeeping activities increased through the years, the General Assembly in 1975 passed a resolution that assessed each member an amount for peacekeeping based on a formula that considers national income, Security Council membership, and other factors. Some nations (including the United States) have neglected to pay their assessments at times, but the system has generated considerably more revenue than one could have expected from the system that it replaced. And because the "optimal" contribution by each nation is likely unknowable, such crude, formulaic mechanisms may do about as well as can be done.

To be sure, such an approach may also fail badly. To the degree that promises to contribute more are enforced by mutual threats to withhold contributions in the event of defection, alliance members may doubt their credibility. For example, members may recognize that in the event of a funding shortfall, those with a substantial private interest in the outcome of a particular conflict may be led to undertake collective defense unilaterally or in a smaller coalition. It may even be possible for would-be free riders to avoid defecting on their promises simply by blocking the alliance from acting in a given conflict, forcing members with high stakes to strike out on their own.

The notion that smaller states will free ride on the efforts of larger states, and the suggestion that U.S. allies in particular free ride on U.S. defense efforts, has long been a theme in strands of the literature on alliances. Convincing tests of the hypothesis are again complicated by the fact that each country's proper share of collective defense in the absence of free riding is difficult to determine at best. Olson and Zeckhauser (1966) posited that military expenditures as a percentage of GDP ought to be approximately constant in the absence of free riding. They then presented evidence of a significant positive correlation between GDP and military expenditures as a percentage of GDP in their dataset, supporting the free riding hypothesis in their view. More recent evidence using the same approach suggests that free riding has diminished with time, especially following the Cold War, as might be expected—nuclear deterrence provided substantial non-rival benefits to allied states, while the more modern concern for small scale conflicts and terrorism has led states to shift their spending toward conventional measures to protect their own territories and nationals.

4.2. The laws of war

The international law of war may be divided roughly into two components: rules on the use of force (*jus ad bellum)* which govern the initiation of hostilities and the response to them, and rules governing the conduct of war (*jus in bello)* such as the customary law and conventions on the treatment of prisoners of war. As with many subjects in this area, very little has been written on either set of rules by economically oriented scholars.

Goldsmith and Posner (2005) address *jus in bello*. They posit that war can be conceptualized as a two-stage game. At the first stage, states are uncertain whether they will go to war. They recognize that war is destructive and that they would be better off if they could reach peaceful settlements that reflect the outcome of future war without the need to bear its costs. But like unions and management, they realize that some destructive conflict may occur. Accordingly, they design rules for the conduct of war with the goal of minimizing the joint losses should it happen. Whatever the outcome of possible war, both sides expect to gain if the number of casualties is held to a minimum along with the damage to their economies. To this end, they adopt rules such as those requiring prisoners to be treated humanely rather than killed, and those prohibiting unnecessarily destructive technologies like chemical and biological weapons.

The rules are enforced by mutual threats of retaliation should one side deviate from them during a war. In some instances, this enforcement structure works. The fact that all sides generally refrained from the use of poison gas during World War II may be a good illustration. Any party could have readily manufactured gas and retaliated in the event of deviation. But it may also break down, as illustrated by the highly variable treatment of prisoners of war in various conflicts through the years. When the number of prisoners is large and the costs of adhering to rules of humane treatment become great, for example, Goldsmith and Posner suggest that deviation from the rules is common. Likewise, smaller powers with limited financial resources are considerably less likely to comply with international rules for the treatment of prisoners.

Posner and Sykes (2005) consider the modern rules on *jus ad bellum*. Under the U.N. Charter, to which virtually all important states are signatories, the use of force is allowed only pursuant to a resolution of the Security Council that authorizes force, or in self-defense. Acceptable cases of self-defense are limited to situations involving an actual or "imminent" attack. In apparent contravention of these principles, the United States has indicated that it will consider the use of preemptive force in self defense under some circumstances. Posner and Sykes focus on the welfare implications of preemptive attacks in self defense, suggesting that preemption can be analyzed within a real options framework. They offer a simple two-period model in which a potential aggressor state has private information about its future plans to attack another more powerful state, and cannot reveal it credibly. In period one, the threatened state can choose to attack preemptively or wait. A preemptive attack will eliminate the threat from the potential aggressor, but may do so unnecessarily if the potential aggressor would never have attacked anyway. Absent a preemptive attack, the potential aggressor state will reveal its intentions by attacking or not in period two, and if attacked the threatened state will respond with force. The analysis assumes that the threat of retaliation in period two is inadequate to deter attack, perhaps because the potential aggressor does not care about retaliation (perhaps its leaders expect to escape, or are undeterrable zealots). The model suggests that if the costs of conflict in period two are sufficiently greater than in period one (as, for example, when the potential aggressor state develops powerful new weapons for period two), and if the probability that the potential aggressor will turn into an actual aggressor is sufficiently high, then preemptive attack may be both privately and socially optimal—the expected discounted costs of conflict computed under either welfare criterion may be lower with preemptive attack. The threatened state will only use preemptive force in accordance with its private interest, however, which may lead to too many or too few preemptive attacks. The latter possibility arises because collateral damage to foreign innocents can be much larger if the threatened state waits until period two to take action. In this simple framework, the traditional rule against preemptive measures appears questionable.

Several objections to this conclusion may be offered. First, nothing in the model distinguishes "good guys" from "bad guys." The analytics are the same whether the state threatened with attack is the United States and the potential aggressor state is in league with Al Qaeda, or the threatened state is Stalinist Russia and the potential aggressor is a repressed ethnic state forced into the Stalinist orbit. Indeed, repressive regimes may face growing threats against them more often than liberal regimes. A rule permitting the use of preemptive force may then countenance behavior that is considered undesirable on average, for reasons outside the model. Second, if the danger to the more powerful state is not imminent but a preemptive attack is genuinely desirable, perhaps the threatened state has the time and should be able to persuade the Security Council to authorize preemptive force. The vote of the Council then serves the function of identifying the "good guys" and allowing them to proceed. A difficulty with this argument, to be sure, is that members of the Security Council may vote (or exercise their veto) on the basis of their private interest in the potential conflict rather than the social interest. Threats may be highly asymmetric in their potential effects on U.N. members, and their private interests may be at odds for other reasons as well. It is an empirical question whether the repeat play nature of Security Council interaction can induce cooperation on optimal policies, but many U.N. observers clearly believe that the answer is no. Finally, one might defend the traditional imminence requirement under international law on the grounds that the conditions under which preemptive force is optimal are difficult to verify, and if nations have a right to use it they may invoke that right opportunistically. If preemptive attack is usually undesirable from a social standpoint, therefore, a bright line rule against it may be the best option.

5. International trade law

I now turn to the subject that has received by far the most attention in the existing law and economics literature. As noted previously, my emphasis will be on topics and ideas that suggest broader lessons for the study of international law, although some familiarity with the traditional economic literature on trade is essential as background to what follows.

5.1. Trade policy, trade externalities and trade agreements

The normative economics of international trade suggests that government intervention in trade flows generally reduces global welfare—one of the few propositions on which most economists will agree. But government intervention in trade has been extensive throughout modern history, raising a puzzle as to why and motivating a vast economic literature on the positive economics of trade policy. Early work focused on the incentives facing "large" countries that act as national welfare maximizers, understanding "large" to mean simply that the demand for the nation's exports, or the supply of its imports, is less than perfectly elastic. Such countries have monopoly power over their exports, and/or monopsony power over their imports. Competitive export industries and small consumers of imports cannot organize to exploit that power, but the government can exploit it through taxation. At appropriate import and export tax rates, national welfare will rise because some of the tax revenue is extracted from foreign exporters and consumers. Tariffs set on this basis are sometimes termed "optimal tariffs," and their contribution to national welfare arises through their effects on the "terms of trade"—the ratio of the price of a nation's exports to the price of its imports. Optimal export taxes improve the terms of trade by increasing export prices (inclusive of taxes) relative to import prices; optimal tariffs improve the terms of trade by lowering import prices (net of tariffs) relative to export prices. A classic early treatment is that of Johnson (1953).

Economists quickly realized that these simple models of national income maximization did not explain the pattern of trade intervention very well in practice. Among other things, the pattern of tariffs across industries did not seem closely related to the market power of the importing nation. And on the export side, nations seemed more likely to engage in policies such as export *subsidization* than export taxation, a behavior that cannot reflect national welfare maximization in competitive markets.[4] Economists began to examine different models of what governments maximize, generally drawing on ideas from the contemporaneously burgeoning literature on public choice. A wide range of modeling strategies was employed, all capturing in one way or another the idea that governments behave as though they care about the distribution of national income as well as its magnitude, usually because some interest groups are better organized than others and will reward their leaders more generously for pursuing policies that benefit them. A good illustration of this approach is the model developed by Grossman and Helpman (1995a). They posit that some domestic industries are organized and some not, and that the organized industries commit to schedules of campaign contributions which are a function of trade policy choices by their government. The government maximand is a weighted average of campaign contributions and per capita economic welfare. The model has a wide range of implications, and makes progress with respect to the apparent deficiencies of the national welfare maximization models. For example, when governments behave as Nash actors, equilibrium trade policy involves higher tariffs for well-organized import competing industries than national income maximization would predict. It also involves lower export taxes on organized export industries, even to the point that export subsidies may arise. Rodrik (1995) provides a valuable survey of the work in this area.

Although the primary focus in much of the literature is on domestic policy formulation, many of the models have immediate implications for the role of international cooperation. The general equilibrium tradition in trade theory necessitates the presence of at least two countries in most models, and the noncooperative equilibrium of the models generally reveals externalities. The externality almost always arises because of the terms of trade effects noted earlier—part of the cost of trade intervention is borne by foreign producers and consumers. The failure of governments acting unilaterally to account for the external costs and benefits of changes in the terms of trade leads to equilibrium

[4] Under conditions of imperfect competition, however, export subsidies may emerge from the pursuit of national welfare by individual states, as suggested by the considerable literature on "strategic trade policy." For a survey see Brander (1995).

trade policies that are off the Pareto frontier. As in most other areas of international law, therefore, the role of international agreements is to overcome the externality problem.[5]

Some commentators are skeptical of the notion that terms of trade effects are the engine of international cooperation on trade policy. This skepticism arises in part because trade policy officials rarely discuss them explicitly, and may not even understand them as a reflection of each state's market power. Bagwell and Staiger (1999, 2002) argue forcefully, however, that terms of trade externalities can and should be interpreted as the basis for the "market access" demands that pervade and motivate modern trade negotiations. Exporters seek negotiated reductions in barriers to trade precisely because those barriers affect their net prices on world markets.

Having identified the likely source of externalities, the next question is how trade agreements will address them. If governments are national welfare maximizers, then an optimal trade agreement should maximize global welfare conventionally defined. Under competitive conditions, free trade will do so, although free trade may not lie in the core and hence some transfer mechanism may be necessary. The allocative effects of free trade can be achieved with transfers in the background if the nations that are offering transfers subsidize their exports and the nations that are receiving transfers impose an equal and offsetting tax on them.

Of course, actual trade agreements do not achieve completely free trade (or its allocative equivalent with transfers). An adequate theory of trade agreements must thus explain how they deviate from that benchmark and why. Grossman and Helpman (2002), Bagwell and Staiger (2002), and numerous others offer political economy models in which governments care about the distributional consequences of trade policy, and thus in which the equilibrium trade agreement does not achieve free trade. The common thread is that cooperation will seek to eliminate the terms of trade externalities across nations, but will at the same time allow governments to respect their preferences over distribution. Well organized import competing industries will retain protection in these models, for example, especially if the foreign exporters with whom they compete are poorly organized.

In sum, mainstream international economics, including many of its most prominent practitioners, has developed an impressive set of tools for modeling the externalities

[5] I note briefly two strands of literature in which the terms of trade externality is not the dominant reason for trade agreements. Ethier (2000, 2004) is generally critical of the terms of trade theory of trade agreements, and suggests that countries can benefit from trade agreements due to some "political externality" that is distinct from any terms of trade effects. A somewhat more substantial strand of literature suggests that trade agreements are not motivated by international externalities at all, but by dynamic consistency problems in domestic politics. In these models, governments are imagined to have some preferences over investments in various sectors, but producers know that once investments are sunk, government may change policy in a way that undermines the returns to investment. In anticipation of this behavior, producers will invest in such a way as to minimize their exposure to it, distorting investment away from the preferred pattern. Government may thus benefit from commitment devices, and it is conceivable that a commitment to free (or freer) trade might be valuable given appropriate governmental "preferences." See, e.g., Maggi and Rodriquez-Clare (1998).

that arise when nations act noncooperatively. These models offer a number of predictions that are broadly consistent with actual trade agreements. They suggest that terms of trade externalities lead tariffs to be excessive relative to those on the Pareto frontier, and that cooperation will thus tend to reduce tariffs in general. They suggest that governments motivated by distributional considerations as well as aggregate welfare will not achieve free trade through trade agreements, but instead will tend toward agreements that protect and/or subsidize the interest groups that are better organized (or whose welfare counts more in the view of their governments, for whatever reason). They are also capable of explaining in broad brush terms the evolution of trade agreements over time, particularly the gradual reduction of trade barriers through a series of negotiating rounds (intuitively, organized resistance to further liberalization diminishes after each round of liberalization as industry specific capital in import-competing industries depreciates without replacement). See Staiger (1995b). I will not dwell on these general points any further, nor will I include formal treatments of the underlying models given the readily accessible surveys elsewhere (especially in the *Handbook of International Economics*.) Instead, I turn to work on some of the particular legal issues that arise in the trade area and that bear on some of the more general themes developed in section 3.

5.2. The legal architecture of world trade and its lessons for international law

Formal trade agreements arose in the 19th century, but I focus here on the modern WTO/GATT system. GATT began in 1947, and was limited to trade in goods. The centerpiece of the GATT was its reciprocal commitments to lower tariffs embodied in Article II, along with a general commitment to nondiscrimination among trading partners contained in Article I. The remainder of GATT served three primary functions: it constrained policy instruments that were substitutes for tariffs; it provided for various adjustments to the bargain including tariff renegotiation, amendments, waiver, accessions and preferential trade exemptions to the MFN obligation; and it provided a dispute resolution system.

GATT subsequently evolved through a series of negotiating "rounds," which initially involved little more than further reciprocal tariff reductions. By the time of the Tokyo Round in the 1970's, however, GATT signatories had become increasingly concerned about the growth of various non-tariff barriers that were inadequately disciplined by the original GATT. The result was several plurilateral "codes" on such matters as subsidies, antidumping measures, product standards, and government procurement. The next major stage in the evolution of GATT came in the Uruguay Round, which lead to the creation of the WTO in 1995. The WTO replaced the GATT (although its treaty text was incorporated into WTO law as "GATT 1994"): it elaborated and tightened obligations with respect to non-tariff barriers and made them binding on all members; it created the General Agreement on Trade in Services (GATS), for the first time bringing services trade under multilateral discipline; it created the Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPs), which required all members to create and enforce intellectual property rights in a number of areas; and it overhauled the dispute resolution system.

It is, of course, difficult to know what would have happened if the GATT system had never evolved, but conventional wisdom holds that it has been quite successful.[6] Tariffs have declined steadily from an average rate of around 40% when GATT was founded to an average rate of about 4% presently. And although the history of the WTO/GATT system is not free of tension, it has so far avoided the problem of large scale trade wars that plagued the global economy before it, such as the widespread tariff increases during the 1930's initiated by the Smoot-Hawley tariff in the United States. Member states also appear to comply with the rules of the new dispute resolution system in many though by no means all cases. In a clear majority of cases where a measure has been held to violate WTO law, members have abolished it. And although members have declined to abolish a few measures that have been ruled illegal, they have accepted the measured retaliation authorized by the system as an alternative.

However its success is measured, the WTO/GATT system is surely among the most intricate and complex international legal arrangements. Careful study of the system is important not only for trade scholars, but for anyone interested in the mechanisms of international cooperation.

5.2.1. Coping with complexity: identifying and structuring trade protection

Tariffs were the principal instrument for the protection of import-competing industries historically, but many other instruments may be used for that purpose. Quantitative restrictions on imports are an obvious alternative to tariffs, and other available devices include discriminatory domestic taxation, subsidies, regulatory policies that burden imports disproportionately, state-franchised monopolies that disfavor imports in their purchasing decisions, and discriminatory government procurement policies. Effective agreements to reduce trade protection must attend to this array of protectionist options. The challenge of formulating appropriate constraints on all of these instruments is complicated by the fact that many of them also address "legitimate," non-protectionist objectives. Subsidies may correct other externality problems; regulations may address important issues of safety and quality; procurement preferences may affect national security; and so on.

These observations suggest that an agreement such as GATT must pursue a number of related objectives concurrently. First, signatories must ensure that commitments to reduce protection associated with one policy instrument are not undermined by the substitution of protection using some alternative instrument. Second, and closely related, signatories must manage the challenge of negotiating reductions in protection given all of the available instruments of protection, and will benefit from strategies that reduce the costs of negotiation. Third, to the extent that some instruments of protection produce greater deadweight costs than others, signatories will benefit if they can channel

[6] For a contrarian perspective, see Rose (2004).

whatever protection remains under their agreement into the instruments that do the least damage. Finally, signatories must design agreements that constrain trade protection appropriately on the one hand, while leaving themselves free to pursue their domestic regulatory agendas on the other.

Regarding the first three of these objectives, the central strategy in GATT has been a process of "tariffication"—the conversion of most protectionist measures into tariffs. This process is accomplished through a number of legal measures, such as the general prohibition on quantitative restrictions, import licensing schemes and the like in Article XI of GATT 1994.

Bagwell and Sykes (2004) address the benefits of tariffication. They argue that it lowers the transactions costs of reciprocal trade negotiations directly by reducing the number of protectionist instruments that are part of the negotiation. It may afford additional benefits by reducing uncertainty about the degree of market access that trade concessions afford, an important consideration if trade negotiators are risk averse. Tariffs also have the virtue that when applied in non-discriminatory fashion, they avoid the loss of joint surplus that results from trade diversion (discussed further below). Other instruments, such as quotas, may result in greater trade diversion depending on how they are administered or in other kinds of deadweight loss (see the discussion of regulatory protection below). Finally, because tariffs tend to be more transparent, cheating on tariff commitments is easier to detect than cheating with respect to commitments on other protectionist instruments. As long as the general commitment to tariffication is itself enforceable, therefore, tariffication discourages cheating and facilitates cooperation.

To say that it is valuable to channel protection into tariffs, however, begs the question of what measures constitute protection. Perhaps nowhere is this problem more acute than in the case of domestic regulatory policies. For example, one quite famous and thorny WTO dispute has centered around the question whether a European prohibition on hormone-raised beef imports, ostensibly for health reasons, is really a pretense for measures to protect the European beef industry.

The modeling framework developed by Bagwell and Staiger (2002) suggests a clever solution to such issues, based on the idea that trade agreements should induce nations to "internalize the externality." The externality arises because choices by individual governments affect the prices confronting foreign buyers and sellers on world markets. If nations commit through trade agreements to maintain a particular level of market access—a particular terms of trade with other nations—they can then be allowed freedom to vary their domestic regulatory policies as they wish as long as any effects on the terms of trade are offset by countervailing changes in their tariff and subsidy policies. Through such a mechanism, other nations are insulated from the terms of trade consequences of domestic regulatory policies, the externality is "internalized," and nations will then behave efficiently in relation to their internal welfare judgments. Bagwell and Staiger model this approach explicitly for labor and environmental standards and competition policy, but the strategy has general applicability as long as the only international externality is transmitted through the terms of trade.

In practice, however, trade agreements do not undertake to specify the terms of trade. Such undertakings may seem realistic in the static models of trade theory, but much less so in an environment of volatile exchange rates and innumerable other macro and microeconomic shocks to world prices. The WTO/GATT system does provide compensation for trade injury caused by certain measures, such as the withdrawal of tariff concessions and the introduction of unanticipated subsidy programs (each discussed further below), but it does not grant compensation for many other policies that affect the terms of trade (except through a little-used mechanism relating to "nonviolation nullification or impairment", see Bagwell, Mavroidis, and Staiger, 2002). Instead, a key strategy of the system has always been to constrain protection *per se*, while leaving signatories free to pursue non-protectionist objectives such as health and safety regulation, environmental protection, and national security. But this strategy requires an additional set of rules to distinguish protectionist measures from non-protectionist measures.

The stakes in drawing this line can be considerable. Sykes (1995, 1999) compares regulatory protection to tariffs. Suppose, for concreteness, that an importing nation wishes to protect its grain industry, and that it has a target for the domestic grain price. Assume that this target price can be achieved with a tariff of τ per bushel of grain, which will limit grain imports to the quantity B, and result in tariff revenue to the importing nation of τB. Alternatively, the importing nation can enact a regulation that requires foreign grain producers to incur additional costs of production (for example, the cost of testing all export shipments for the presence of some ostensibly dangerous pest or chemical). Let the additional cost per bushel of grain exports be τ under this regulation, and assume that no tariff will be imposed in this alternate scenario. Assume further that the ostensible health or safety basis for the regulation is pretense, and that the sole motivation for the regulation is to achieve protection. From the standpoint of the grain industry in the importing nation, both approaches are equivalent in their ability to achieve the target domestic price. Likewise, both measures will have the same effect on the terms of trade, extracting surplus from grain exporters if the import supply curve is less than perfectly elastic. But the two approaches are not equivalent in welfare terms—under the regulatory approach to protection, the tariff revenue "rectangle" τB is transformed into pure deadweight loss (given the absence of any bona fide health or safety justification). Wasteful regulations of this sort may hold no appeal to importing nations that are unconstrained in their tariff policies, for they would prefer to achieve the target level of protection *and* to enjoy the tariff revenue. But if tariffs have been constrained through reciprocal trade negotiations, the temptation to engage in wasteful regulatory protection can emerge. Of course, if the importing nation cares somewhat about its national income, regulatory protection will be less attractive than tariff protection, and one would expect only a partial substitution between regulatory measures and tariff measures. The important point is that if trading nations can substitute regulatory protection without penalty under the law, they will be tempted to do so to some extent. And the Nash equilibrium of the game in which nations make these substitutions can entail considerable welfare costs. It is in the mutual interest of trading nations to prohibit such measures—it is easy to see that by comparison to any situation with regulatory protection, affected importing and exporting nations could eliminate it, adjust their tariffs in some fashion, and make themselves all better off.

With particular reference to domestic regulatory policy, the initial response of GATT to this problem was to prohibit discriminatory regulations through a "national treatment" obligation in Article III. No regulatory burden could be imposed on imports unless it was also imposed on domestic producers of "like products." GATT also provided a number of exceptions to this principle in Articles XX and XXI, which cover things such as measures "necessary" to protect the environment or measures "considered essential" to national security. To the degree that the national treatment obligation was limited to facially discriminatory regulations, however, it soon proved inadequate. Consider a "non-discriminatory" regulation that requires all products to embody a certain design or technology, when domestic producers have a cost advantage in producing or obtaining that design or technology, and when some other design or technology available more cheaply abroad will achieve the non-protectionist regulatory goal just as effectively. The cost difference between the two technologies then represents pure regulatory protection. The need for additional disciplines to deal with these and related issues culminated in two important new agreements during the Uruguay Round, the Agreement on Technical Barriers to Trade and the Agreement on the Application of Sanitary and Phytosanitary Measures (pertaining mainly to pests, disease-causing organisms and contaminants in foodstuffs). These agreements impose numerous additional constraints on domestic regulators, such as general least restrictive means requirements and a requirement that product regulations specify performance goals rather than design limitations. Somewhat more controversially, they also introduce requirements that certain health and safety regulations have adequate scientific basis. To return to the example of the European prohibition on hormone-fed beef imports, that regulation was found to violate WTO law because it lacked adequate scientific foundation. For further discussion see Trebilcock and Soloway (2002).

Where domestic regulations appear to have bona fide, non-protectionist goals, however, the primary constraint on them remains a simple national treatment requirement. Such a rule tolerates changes in domestic regulatory policy even if they have significant trade externalities, but it is not clear that any alternative rule is better. Battigalli and Maggi (2003) develop a model of international agreements on product standards, in which negotiators cannot anticipate proper standards for future products. The model suggests that only a national treatment requirement is joint welfare enhancing *ex ante* for arbitrary distributions of future products.

Another important body of WTO law concerns the use of subsidies. Most of the work that has been done on this subject is normative, asking whether WTO rules on subsidies promote global economic welfare conventionally defined. Early writers observed that subsidies to domestic industries can protect them just as effectively as tariffs, yet subsidies may also be justified for reasons other than protectionism. A question then arises whether rules can be developed to distinguish "good" subsidies from "bad" subsidies, avoiding the welfare loss from the former while preserving the welfare gains from the latter. Schwartz and Harper (1972) are deeply skeptical of this enterprise, arguing that workable rules for the identification of welfare-reducing subsidies are impractical, particularly if one allows that "legitimate" subsidies may include those where the citizenry is willing to pay to preserve certain forms of inefficient enterprise (such as cultural industries or family farms).

Nevertheless, the WTO system has evolved increasingly detailed rules governing subsidies. In goods markets, the governing principles are three: (1) New and unanticipated subsidy programs that upset market access expectations are illegal, even if they do not violate any provision of the treaty text, under the rubric of "nonviolation nullification or impairment"; (2) export subsidies (subsidies that favor exportation over domestic production) are illegal (putting aside some exceptions in the agriculture sector); and (3) domestic subsidies, which are identified by the criterion that they must be targeted to a single industry or narrow group of industries, may be illegal if they cause "injury" to foreign exporters or import competing industries, and may also be a basis for unilateral countervailing duties (duties to offset the value of the subsidy) by importing nations whose industries are injured by subsidized imports. Sykes (2005a) evaluates these rules from the traditional welfare economics perspective. Putting aside the difficult question of what constitutes a "subsidy," the first principle seems defensible as a way to protect the value of the bargain associated with tariff concessions against opportunistic erosion. The second principle may also make sense from the perspective of traditional welfare economics, as it is difficult to imagine any constructive role for export subsidies as a first-best policy. But the third principle is problematic in that it defines "subsidies" using a dubious criterion—the targeting criterion may well condemn useful subsidies, and is quite underinclusive as to economically wasteful subsidies or as to subsidies that may have harmful external effects. Further, regardless of the degree of targeting, it is not clear that coherent criteria can be developed for determining the existence of domestic subsidies at all given the vast range of tax/subsidy and regulatory policies that affect the competitive position of firms—WTO rules view each government program in isolation, and make no effort to assess the net impact of government as a whole on the competitive position of firms.

Bagwell and Staiger (2006) also express skepticism toward recent enhancements in the WTO rules on domestic subsidies. They emphasize the possible utility of subsidies to the pursuit of desired efficiency and distributional objectives by WTO member states, and suggest that recent restrictions on subsidies may interfere with their use for desired domestic purposes. If so, and if the effect is important enough, WTO members may be discouraged from negotiating market access commitments to the degree that such commitments increase the likelihood of a challenge to their subsidy policies.

Finally, as developed at length in Sykes (1989), the unilateral use of countervailing duties by importing nations is exceedingly difficult to defend if the policy objective is to maximize either national or global economic welfare (conventionally defined). Countervailing duties almost always will lower the welfare of nations that use them, unless by coincidence they are applied in a manner that exploits the monopsony power of the importing nation (recall the "optimal tariff"), or are applied in industries where domestic firms generate positive externalities (offering a rationale for "strategic trade policy"). From the global perspective, the uncoordinated use of countervailing duties by individual trading nations is unlikely to do much to discourage wasteful subsidy practices, especially since the criteria for identifying and measuring countervailable subsidies suffer from the deficiencies noted above.

Moving from the normative to the positive, WTO rules on subsidies raise a number of puzzles to which satisfactory answers do not yet exist in the literature. Bagwell and Staiger (2002) focus on export subsidies. Following the political economy literature on trade policy, they posit that governments may give considerable weight to the welfare of exporting industries. Export subsidies may then become optimal from a political standpoint. They consider a three-country model with two exporting nations and one importing nation. Absent cooperation, the two exporting nations may find themselves in a "subsidies war"—in Nash equilibrium their export subsidies impose a negative externality on each other (diverting exports from one to the other) and the private benefits of subsidization exceed the joint benefits. If the two exporting nations then cooperate, they will strike a deal that reduces the degree of subsidization and, depending on parameters, it is possible that such a deal would eliminate export subsidization altogether. Now consider adding the importing nation to the bargain. In the Bagwell/Staiger framework, the importing nation will benefit from export subsidies, *ceteris paribus*, which improve its terms of trade. Thus, unless the importing nation attaches substantial weight to the welfare of the import-competing industry, an agreement involving all three nations might well lead to greater export subsidization than would be observed if only the exporting nations were to cooperate. It is by no means obvious, therefore, why a multilateral agreement should be hostile to export subsidies across the board. More generally, in the absence of international cooperation, the general tendency is for unilateral trade policy to result in too little trade. Because export subsidies increase the volume of trade, it is something of a puzzle as to why multilateral cooperation systematically condemns them.

I conclude with a note about services trade, which has been subject to WTO rules (GATS) only since 1995. Conceptually, the externality in the services area is the same as in the goods area—restrictions on services imports can extract surplus from foreign service providers. The mechanism that gives rise to this externality will often be different in the services area, however, as will be the approach to ameliorating it. The reason lies in the fact that tariffs are generally not possible with services trade because services imports rarely cross the border in a manner that allows them to be taxed. Instead, most of the market access issues in the services area involve efforts by foreign service providers—banks, insurance companies, law firms—to establish a physical presence in the importing nation. The obstacles to the entry of foreign service providers are then regulatory, in the form of licensing requirements, prudential requirements, residency requirements, immigration restrictions, and the like. Some such regulations are applied in a facially discriminatory fashion, and in other instances regulation simply imposes a disproportionate burden on foreign service suppliers. See Trebilcock and Howse (1995). The liberalization of trade in services thus poses difficult new challenges. The process of "tariffication" that characterizes the approach of GATT to goods markets simply cannot work here, and other strategies must be employed to reduce the number of protectionist instruments. Further, tight nondiscrimination requirements such as those seen in GATT might eliminate the ability of GATS signatories to protect their service industries at all, a result that is unlikely to be politically palatable. Hence, the strategy of GATS is to allow discriminatory policies to be "scheduled:" In each service sector, signatories make a determination whether to afford national treatment or not to foreign service providers. If they do so, they can nevertheless reserve the right to discriminate in particular ways by listing the discriminatory policy in their schedule of service commitments. Although this approach enables nations to retain the protection that they regard as essential and makes their trade barriers more transparent, it does little by itself to channel protection into the least wasteful policy instruments. On that front, GATS has much more to accomplish. See Sykes (2001).

5.2.2. Multilateralism, bilateralism, regionalism and trade discrimination

The world trading system exhibits a multiplicity of trade agreements, some bilateral (such as the U.S.–Israel Free Trade Agreement), some regional (such as NAFTA and the EU) and some global (the WTO). What explains this pattern? The answers to this question are not fully satisfactory even though the economic literature in the area has grown enormously in the last fifteen years. In this section I can only hope to identify some highlights.

Perhaps the first question to ask is why do nations go beyond bilateral agreements if cooperation is generally harder to orchestrate as the number of parties to an agreement increases? The answer, not surprisingly, is that bilateral agreements create important externalities. Imagine three countries, A, B, and C, each of which trades with the others. Imagine further that tariffs on imports into each country are initially nondiscriminatory. Suppose that A and B contemplate a trade agreement in which each lowers its tariffs on imports from the other, and suppose for concreteness that A lowers its tariff on imports of grain from B. The effect on C will depend on whether it is an importer or exporter of grain. If C initially exports grain to A, then the tariff reduction on A's imports from B will cause imports into A from B to rise at the expense of imports from C. The weakening in demand for C's exports causes C's terms of trade to worsen. And to the degree that the tariff preference for B causes consumers in A to shift purchases from C to B, the phenomenon of *trade diversion* arises. In general, trade diversion occurs whenever discriminatory trade policies cause consumers in an importing nation to make purchases from abroad that would not be made but for the discrimination. It yields a deadweight loss in global welfare, *ceteris paribus*, because imports do not originate from the lowest cost supplier. Viner (1950) is a classic reference. Another type of externality arises if C is instead an importer of grain. The preference for grain imports into A from B will then tend to increase the cost of grain to C. C's terms of trade worsen because the price of its imports increases.

Of course, an agreement to liberalize trade within a subset of trading nations also has its benefits. Most importantly, if an agreement leads to an expansion of trade between nations that are efficient suppliers of each others' markets, the phenomenon of *trade creation* arises. It is entirely possible that the benefits of trade creation will exceed the losses from trade diversion, so that a bilateral or regional agreement can on balance enhance global welfare as conventionally defined, and can assuredly enhance the welfare of its members.

Because of the terms of trade externalities for non-members, however, any agreement involving only a subset of trading nations has the potential to injure non-parties to the agreement. Formal models of 3 or more country trading networks thus tend to suggest that agreements among any subset of countries will lead to inefficiencies in Nash equilibrium unless some constraint is imposed on the terms of those agreements. In that regard, Kemp and Wan (1976) derive an important result about the formation of agreements known as *customs unions*, in which every party to the agreement harmonizes its external tariff. They show that such an entity can adjust its common external tariff so that the terms of trade with non-members remain the same as before the formation of the union. The welfare of non-members is then unaffected, and the gains to the member states are equal to the global gains. Their result is closely related to the concept of reciprocity developed in Bagwell and Staiger (2002).

In light of the mischief that may result from discrimination in international trade, however, one wonders whether a prohibition on discrimination may be valuable. Precisely such a general prohibition is contained in Articles I and XIII of GATT, which prohibit discriminatory tariffs and quota regimes (when the latter are allowable at all). This obligation is termed the most-favored-nation (MFN) obligation. By joining a multilateral agreement such as GATT and committing themselves to respect the MFN obligation (subject to a very important exception to be discussed in a moment), signatories avoid the loss of joint surplus associated with trade diversion that could arise in a smaller agreement. And in addition to ensuring that goods are imported from the lowest cost foreign supplier, the MFN obligation may facilitate trade negotiations in an important way. Imagine once again a trading system consisting of countries A, B and C, and suppose that A and B contemplate a trade agreement with each other. Negotiators for A and B may worry that after concluding an agreement, their partner may later negotiate an agreement with C, offering C a more attractive arrangement. Whatever benefits they expect to get from the agreement may be undermined, and they may be reluctant to conclude any agreement at all. But if they mutually promise each other that no third nation will be offered better terms later through an MFN commitment, such worries can be put to rest. Schwartz and Sykes (1996).

To be sure, the MFN obligation may also carry a cost. If benefits extended by A to B must be extended automatically to C because C is entitled to MFN treatment, then C may become a free rider on trade negotiations between A and B. To solve the free rider problem, it may be necessary for A and B to draw C into the negotiations. And as the number of countries drawn into negotiations rises, the costs of negotiation rise and the problem of holdouts surfaces.

Thus, the role of the MFN obligation in trade agreements raises cross-cutting and complex issues. A considerable literature now exists on the subject, which confirms that the welfare implications of the MFN obligation are complicated, to say the least. A valuable recent survey is that of Horn and Mavroidis (2001).

The reader may note, however, that even if an MFN obligation is on balance desirable in the trading system, that observation by itself does not necessitate multilateral agreements. One could imagine in the abstract a web of bilateral agreements, each containing an MFN clause. We then return to the question posed at the outset—what is the value of multilateral agreements in this context?

A trivial answer is that economies of scale may exist to the degree that the optimal bilateral agreements would all contain many of the same terms. It may then be easier to write a single multilateral agreement. Further, given the free rider problem that arises if bilateral agreements contain an MFN obligation, it may make sense for all trading nations to meet and negotiate at once, each holding off any final deals until all offers are on the table. Bagwell and Staiger (2002, 2004) also show that even when an MFN obligation precludes the possibility of discriminatory tariffs, bilateral negotiations nevertheless create terms of trade externalities for third countries. For example, if nation A offers a tariff concession on some product to nation B and extends it to all other exporters of the product pursuant to a most favored nation obligation, a worsening of the terms of trade will still occur for any third nation that imports the product. Such externalities provide further justification for multilateral negotiations.

Maggi (1999) identifies another potentially important consideration favoring multilateral cooperation. Imagine three countries once again, A, B and C, and suppose further that A runs a large trade surplus with B, B runs a large surplus with C, and C runs a large surplus with A (although aggregate trade for each country is balanced). In this scenario, an imbalance of power may be said to exist in each bilateral relationship. Because of that imbalance, the amount of self-enforcing cooperation that is sustainable in bilateral tariff agreements is limited—if B defects from an agreement with A, for example, A loses much more than B loses if A defects from the agreement. This scenario suggests a role for an agreement with a multilateral enforcement mechanism, whereby defection by any party can be punished by all of the others. As Maggi acknowledges, however, the extent to which such coordinated enforcement occurs in practice within the WTO is quite limited.

If trade discrimination is often problematic, and if important externalities argue for multilateral trade negotiations, an important puzzle remains. GATT Article XXIV permits signatories to form customs unions and free trade areas (the latter differing from a customs union in that the members do not harmonize their external tariffs). Both customs unions and free trade areas permit member nations to deviate from the most favored nation obligation, as long as they eliminate barriers on "substantially all" trade between them. Why should GATT include this gaping exception to the most favored nation obligation, allowing discrimination as long as it occurs on a grand scale ("substantially all trade")? Related, is the proliferation of preferential trading arrangements such as the EU, NAFTA, and Mercosur a positive or unfortunate development for the international trading system?

Note that nothing in Article XXIV ensures compensation to nations that are injured in their terms of trade by the formation of a preferential trading arrangement.[7] For this reason, an enormous theoretical and empirical literature has developed on preferential trading arrangements, and I will make no attempt at a survey here. In general, theoretical work tends to be skeptical of preferential trade for the reasons suggested by the discussion above. Valuable windows into the literature include Bagwell and Staiger (2002), Bhagwati and Panagariya (1996) and Grossman and Helpman (1995b). Empirical work perhaps tends to be more agnostic, with some support for the idea that the beneficial trade creation under existing agreements may exceed the harmful trade diversion. But on the ultimate question whether preferential arrangements are "building blocks" or "stumbling blocks" to multilateral cooperation in the words of Lawrence (1996), the issue remains unsettled.

One other form of discrimination in the WTO system has received a bit of attention. Pursuant to the "enabling clause," developed country members of the WTO are permitted to offer tariff preferences to exports from developing countries, ostensibly to promote development. A number of nations do so pursuant to schemes that are commonly termed the "Generalized System of Preferences" (GSP). These schemes by their design discriminate between developed and developing country exports. Many of them also discriminate across developing countries, awarding or denying preferences based on differences in per capita GDP, differences in relative success of a particular nation at exporting a particular good, and various political considerations. India brought a legal challenge to the European scheme in the WTO dispute resolution process, urging that the criteria it used to discriminate among developing countries were impermissible. Grossman and Sykes (2005) consider the economics of trade discrimination in favor of and among developing countries, both from a theoretical and empirical standpoint. They question the wisdom of discrimination in the existing schemes, and suggest that GSP schemes likely do little to promote economic development.

5.2.3. *Enforcing trade agreements*

I begin with a review of the WTO/GATT dispute resolution system, a review that will highlight a number of puzzles. Shortly after GATT was formed, the organization gravitated to a "consensus" rule for many decisions. Under the consensus rule, an investigation into an alleged violation of the agreement (undertaken by a dispute "panel") could only be authorized by consensus (unanimity). Even if the panel was constituted

[7] If the arrangement is a customs union, Article XXIV does require compensation if the new common external tariff violates a tariff commitment previously made by a member state (e.g., if after Turkey joins the EU, it raises its tariff on widgets above the level it had previously promised to some non-EU trading partner, the EU must compensate that partner). But this compensation obligation is incomplete (it does nothing to compensate for the increase in market power enjoyed by the trading block, for example) and its efficacy in practice has proven questionable. Moreover, when the preferential arrangement is a free trade area, members do not alter their tariffs on imports from non-members at all, and no compensation is required.

and ruled in favor of the complaining nation, the ruling still had no force of law unless it was "adopted" by the membership, again requiring consensus. Finally, sanctions for violations following the adoption of a ruling could only be authorized by consensus. Thus, the violator nation could always block investigation, adoption, and sanctions. Nations often allowed investigations to go forward nevertheless, and also in many cases permitted adverse rulings to be adopted. But sanctions were only authorized once during the history of GATT (prior to the creation of the WTO).

Under this system, cheating or alleged cheating could only be punished through unilateral "self-help." Nations such as the United States would bring cases to the GATT, and retaliate unilaterally if the target country refused to comply with an adverse ruling or blocked the process from going forward. Studies of the efficacy of unilateral retaliation during this phase of GATT indicate that its results were mixed, although retaliation often induced nations to alter their policies, and many nations were observed to comply with adverse panel rulings without the need for any retaliation. See Sykes (1992), Bayard and Elliott (1994).

In 1989, GATT signatories agreed to end the ability of accused nations to block an investigation, although they could still block the adoption of panel reports and block sanctions. Finally, with the creation of the WTO in 1995, the system put an end to blocking altogether. Rulings by a dispute panel (or the new Appellate Body) are automatically adopted unless a consensus exists against adoption, and complaining nations are entitled to use sanctions as a matter of right against a party that has been adjudicated to be in violation of WTO law and that refuses to bring its behavior into compliance. Retaliation is limited to the withdrawal of "substantially equivalent" trade concessions, and the level of retaliation under this criterion is subject to arbitration.

The system has one other unusual feature. If a nation violates WTO law, it is not sanctioned for the violation *per se*. Rather, a sanction is permissible only after the nation in question has been adjudicated to be in violation and has been given a "reasonable time" to comply with the adverse ruling. A nation that complies within a "reasonable time" thus faces no sanction at all. Further, although the treaty text is not entirely clear on the matter, the general view is that any sanctions should be "substantially equivalent" to the prospective harm from the refusal of the violator nation to comply going forward—there is no sanction for the trade injury caused by the violation until the "reasonable time" for compliance has elapsed.

How well do the available economic models explain the history and practice of WTO/GATT dispute resolution? Standard models of trade agreements suggest that such agreements are self-enforcing, with cooperation supported by mutual threats of defection. Bagwell and Staiger (2002) provide a good illustration. They imagine that two nations agree to cut tariffs in an infinitely repeated game. Each nation is reasonably patient, and plays the "grim" strategy—if either nation cheats in one period, the other reverts to its Nash equilibrium tariff choice in the next and all future periods. Recurring cooperation is then a subgame perfect equilibrium as long as the cooperative tariff does not create too much incentive to cheat in the short term. This latter requirement places a limit on the amount of tariff cooperation that is self-enforcing, and it is possible that the politically first-best tariff level is unsustainable.

In such models, trade retaliation (as distinguished from compensation for an adjustment to the bargain, discussed below) is generally an out of equilibrium behavior and will not be observed in practice (although retaliation may be observed in a suboptimal equilibrium of an infinitely repeated game). Subgame perfection may also rule out strategies such as "tit for tat," which allow retaliation to end once cooperation has been restored. There is also no role in such models for a formal dispute resolution system. They thus provide quite an incomplete account of WTO/GATT enforcement, where retaliation is in fact observed, is transitory in almost all cases, and is now (generally) seen only after an adverse ruling by a formal dispute settlement proceeding.

Some features of the system may perhaps be explained by the considerations developed above in section 3. The formal dispute resolution system, for example, serves as an information revelation mechanism. WTO members may be unable to discern on their own whether another member has cheated, and a formal investigation may be helpful to identify cheaters and trigger reputational sanctions. In addition, the dispute resolution system may serve as a source of default rules for an incomplete bargain, approximating what WTO members would have negotiated had they addressed the issues directly.

Schwartz and Sykes (2002) address the question of why the old GATT system with "blocking" was replaced by the new WTO system with automatic sanctions. Some commentators suggest that the new system was designed to toughen sanctions and discourage cheating. Schwartz and Sykes argue, however, that the new system was designed to *reduce* rather than increase the penalty for cheating. They observe that the penalty for cheating in both the old and the new systems is essentially the same—aggrieved nations retaliate with trade sanctions. What has changed is that the aggrieved nations can no longer set the level of sanction themselves without central oversight. Instead, all sanctions are subject to arbitration to determine whether they are "substantially equivalent" to the harm caused by the ongoing violation. The new system thus reins in the magnitude of unilateral retaliation. This is valuable if excessive retaliation can destabilize the system by triggering a trade war, or if it is important to calibrate retaliation to facilitate "efficient breach" (discussed further below).

I am unaware of any satisfactory explanation for the fact that sanctions are prospective, and limited to situations in which violator nations refuse to comply with rulings after a "reasonable time." Perhaps litigation in the WTO has large positive externalities in clarifying the terms of the bargain, and the limited prospect of sanctions encourages disputants to litigate to conclusion. Or perhaps reputational concerns alone are enough to discourage most blatant cheating, so that the bulk of cases are expected to involve good faith disputes over legal obligations. Because trade sanctions are costly (creating deadweight costs and perhaps political costs as well), the membership may not wish to sanction behavior that arises in good faith. Sanctions may then be used only as a last resort against recalcitrant cheaters. These suggestions are little more than conjectures, however, and a fully convincing explanation remains to be developed.

One last feature of the system that has drawn some attention is the problem of power asymmetries. It is often suggested that smaller nations have no effective power to retaliate for trade violations. Because they have little market power, trade sanctions simply raise prices to their consumers with little impact on the targets of sanctions. Mavroidis (2000) provides some evidence that smaller nations are indeed less effective in securing redress through the WTO dispute resolution system. In response to this problem, Mexico has proposed within the WTO that retaliation rights be tradable. Bagwell, Mavroidis, and Staiger (2003) formalize this suggestion in an auction model. If the violator nation is allowed to participate in the auction, their model suggests that it will win the auction and retire the retaliation rights. This outcome has the interesting feature that trade retaliation never actually occurs, and is replaced by a monetary payment to the nation that initially has the right to retaliate. Smaller nations thus gain some leverage, and trade retaliation with its deadweight costs is replaced by a monetary sanction that is a pure transfer. It remains to see whether such proposals have any "legs" within the system.

Some related issues about the dispute settlement system remain relatively unexplored. For example, if it is indeed the case that trade sanctions are perceived as costly by WTO members, why does the WTO not embrace an alternative sanctioning regime, such as one involving money damages that are a pure transfer? Guzman (2004) raises the general puzzle of why international agreements use what often seem to be inefficient sanctions, and why money transfers are not used with any regularity. Further, why is standing limited to WTO members, even though violations of WTO law plainly impose substantial costs on private interest groups?

On the choice between trade sanctions and money damages, Sykes (2005b) argues that money damages are not a "pure transfer" given the deadweight costs of taxation, the potential effect of money damages on litigation costs, and the challenges that many developing countries would face in raising the funds to pay money damages. The question of which sanctioning regime is more efficient is perhaps not quite as easy as some commentators suggest. Regarding private standing, Levy and Srinivasan (1996) note that the decision by private actors to bring dispute cases has external effects, such as the possibility of an unfavorable impact on other aspects of international relations. Governments may wish to avoid this problem by denying standing to private parties. Sykes (2005b) makes some related points about private standing that will be noted in section 6 below, which addresses international investment agreements.

5.2.4. Adjusting the bargain

Like most long-term agreements, WTO commitments are negotiated under conditions of considerable uncertainty about the future. The treaty text thus contains many provisions for adjusting the bargain over time. Here I will emphasize three features that have received attention in the literature: provisions for renegotiation; an "escape clause"; and a mechanism for the facilitation of efficient breach.

In the WTO as in many commercial contexts, it is no doubt too costly to write a complete contingent contract. Commitments entered at one point in time may then prove undesirable (at least from the political standpoint of government officials) at some future time. One solution, of course, is simply to renegotiate when circumstances change, recognizing that renegotiation is a potential double-edged sword if sunk costs can be exploited.

Ordinarily, agreements do not require express provisions providing for renegotiation, in as much as renegotiation is always a possibility without them. But GATT included a provision for the renegotiation of tariff commitments, Article XXVIII, in its original 1947 text. Article XXVIII provides that a nation wishing to withdraw a tariff concession should attempt to obtain the permission of affected nations by offering to substitute some other concession. If such negotiations fail to reach agreement, however, a nation can unilaterally withdraw its tariff concession. Affected nations may then retaliate by withdrawing "substantially equivalent" concessions of their own. As Bagwell and Staiger (2002) argue, this structure can be seen as maintaining reciprocity. Roughly speaking, a nation that proposes to withdraw a concession must restore the terms of trade for affected nations, either by substituting an alternative concession or by accepting a withdrawal of concessions on its exports.

But why is it necessary to memorialize this mechanism in Article XXVIII, instead of simply recognizing that GATT members can renegotiate the agreement when they wish? Schwartz and Sykes (2002) offer the suggestion that the members desired to make clear that tariff concessions could be withdrawn if necessary without securing the permission of affected nations beforehand. In a rough sense, they wished to create a "liability rule" rather than a "property rule." The reason for preferring a liability rule here is the holdout problem. If the permission of all affected GATT members had to be obtained before a politically uncomfortable concession could be modified, the negotiation process could drag on indefinitely and the opportunity to adjust the bargain to changing circumstances would be diminished.

Of course, renegotiation is but one option for avoiding the performance of obligations that become inefficient due to changing circumstances. Drawing on the analogy to commercial contracts, such agreements routinely contain provisions that excuse performance under various contingencies—*force majeure* and Act of God clauses are illustrative. In the GATT system, a similar state-contingent device for excusing performance is the Article XIX "escape clause." It provides for the temporary suspension of tariff concessions when, due to unforeseen developments, increased quantities of imports cause or threaten to cause "serious injury" to an import-competing industry. The escape clause thus permits temporary tariff increases, under specified conditions, in response to import surges.

Formal models of the role of the escape clause include Bagwell and Staiger (1990) and Sykes (1991). Bagwell and Staiger (2002) suggest that import surges risk destabilizing cooperation because they enhance the incentive to cheat. Following an import surge, the short-term incentive to cheat to exploit terms of trade gains increases, as may the political incentive to cheat if the import-competing industry is well organized. They suggest that to avoid unraveling of the agreement that might occur if tariff increases under these circumstances were defined as cheating, the GATT agreement instead allows temporary tariff increases during the import surge. Sykes takes a slightly different approach, modeling the escape clause as a state-contingent rule that permits parties to an agreement to withdraw a concession whenever it is no longer jointly optimal for the concession to be honored—essentially, when the cost of performance to one party exceeds the benefit to the other party (in political terms evaluated at appropriate shadow prices). Such a rule increases the expected utility of trade concessions and thereby allows more concessions to be negotiated in the first instance. The model thus lends support to the suggestion by Dam (1970) that the function of the escape clause is to encourage trade concessions *ex ante*.

Considerable controversy has arisen in recent years, however, as to the circumstances under which WTO members may resort to the escape clause. Much of the problem relates to the confusing treaty text that requires, among other things, that increased quantities of imports must cause or threaten serious injury to an import-competing industry. How is this to be interpreted, given that the quantity of imports is an endogenous rather than a causal variable? For a survey of the current controversies and some suggestions by economically oriented scholars regarding their possible resolution, see Grossman and Mavroidis (2004) and Sykes (2006a).

A final mechanism for adjusting the bargain is suggested by scholarly work on contract damages. It is now well known that a rule of "expectation damages," which requires a party who breaches a contract to compensate the other for its losses, induces the breaching party to internalize the joint costs of breach and thus to breach only when it is "efficient." Schwartz and Sykes (2002) suggest that the rules for retaliation under WTO law, which allow aggrieved parties to withdraw "substantially equivalent" concessions and subjects retaliatory withdrawals to arbitration under this criterion, may be roughly equivalent to an expectation damages regime, allowing aggrieved WTO signatories to retaliate to a degree that approximately restores their pre-breach political welfare but no more. The opportunity for signatories to "buy out" their obligations under this rule is particularly valuable in circumstances where the other options for adjusting the bargain, such as tariff renegotiations and the escape clause, do not address the underlying political problem. The beef hormone controversy between the United States and Europe is perhaps illustrative. If one assumes that European officials are under intense pressure to prohibit hormone-raised beef from entering the domestic market, they cannot achieve this limited goal by adjusting most-favored-nation tariffs applicable to all beef in an Article XXVIII renegotiation, or by invoking Article XIX on temporary import surges. The efficient breach hypothesis regarding the role of calibrated retaliation in the WTO receives further exploration and critique in Lawrence (2003).

5.3. Miscellaneous trade issues

I conclude this chapter with two further issues not yet considered, both of which have received considerable attention in law and economics writings.

5.3.1. Antidumping policy

"Dumping" refers to certain pricing practices by private firms engaged in international trade. Although its precise meaning has changed through the years, dumping occurs under modern trade laws when an exporting firm's prices satisfy (roughly) one of three conditions: (a) its F.O.B. price to the complaining export market is below its F.O.B. price to its home market for the same goods (or for similar goods adjusted for cost and quality differences); (b) in the absence of substantial home market sales of identical or similar goods, its F.O.B. price to the complaining export market is below the F.O.B. export price to some third country market; or (c) its F.O.B. price to the complaining export market is below the fully allocated cost of production for the good in question (including an allocation of fixed costs, general selling and administrative expenses, and so on). In some contexts, the term "dumping" is also sued to refer to exports that have been subsidized, but U.S. law has long distinguished between dumping and subsidization and I will maintain the distinction here. WTO rules on subsidies and related countermeasures (countervailing duties) were briefly discussed earlier in this chapter.

GATT Article VI provides that dumping "is to be condemned," but does not prohibit it or impose any obligation on WTO members to prevent or punish it. Instead, importing nations are permitted to take countermeasures against dumping, in the form of an "antidumping duty" that may not exceed the "margin" of dumping found to exist on the goods in question. That margin is equal to the difference between the "fair value" of the goods computed using one of the three benchmarks above, and the F.O.B. export price to the country imposing the duty. For example, under the first criterion for dumping, the margin would be computed as the F.O.B. price to the home market minus the F.O.B. price to the export market. GATT does not permit antidumping duties in all instances of dumping, however, but limits them to cases in which the dumping is causing or threatening to cause "material injury" to the import-competing industry, a requirement known as the "injury test." These requirements were elaborated, along with the procedures for the conduct of antidumping investigations, in the WTO Antidumping Agreement, the successor to the Tokyo Round Antidumping Code.

Antidumping laws first emerged in Canada in the early 1900's, and quickly spread to the United States. Their proponents put them forward as an adjunct to antitrust statutes, plugging a purported "loophole" that would otherwise allow foreign firms to engage in monopolization through aggressive pricing. A moment's reflection on the standards for dumping above, however, suggests that they are radically different from the variable cost-based standards for predatory pricing that have evolved under modern antitrust law. Price discrimination dumping, reflected in possibilities (a) and (b) above, can assuredly occur at prices above variable cost. And possibility (c) above involves not a comparison between price and some measure of variable cost, but a comparison between price and (roughly) a measure of long run average cost. Further, although the material injury test requires that some harm befall import-competing firms as a prerequisite to antidumping duties, it falls far short of a structural analysis of the industry to determine whether monopolization is a plausible outcome, and antidumping duties are routinely observed in industries producing such items as steel products, potatoes, textiles and footwear—all industries that are highly unconcentrated on a global level and where a danger of monopolization is utterly implausible as a basis for antidumping measures.

The connection between antidumping policy and sensible antitrust policy is thus a tenuous one at best. Sykes (1998) reviews the economic and political history of antidumping laws at considerable length, and concludes that their proponents were well aware that antidumping measures were not a necessary part of antitrust policy. Instead, the goal was to provide additional tariff protection to troubled industries plagued by import competition, an objective akin to the rationale behind the GATT escape clause discussed earlier.

Nevertheless, early economic writing on antidumping policy took a favorable view of it. Viner (1923) argued that low prices attributable to dumping are transitory, and asserted that they may impose adjustment costs on the importing nation that exceed the benefits of temporarily cheaper imports. The basis for Viner's analysis, however, was shaky. Dumping need not be transitory at all (the price discrimination form of dumping arises because different markets have different demand elasticities, a fact that may well persist over time). Even when dumping prices are transitory, nothing in Viner's work (or since) demonstrates that temporarily cheap imports systematically impose adjustment costs that exceed the welfare gains to the importing country from temporarily cheaper imports. If actors in import-competing industries have rational expectations, one would expect them to incur adjustment costs only to the extent that they are cost-justified from a social standpoint.

Not surprisingly, therefore, economic commentators eventually began to question the wisdom of antidumping policy. The lack of any connection between antidumping policy and sensible anti-predation measures has now been noted by commentators too numerous to mention. Others have noted that an antidumping duty may prove beneficial to an importing nation under special circumstances, as when it has not yet imposed the "optimal tariff' to exploit its monopsony power, or where it protects a "strategic industry" with positive spillovers, but in such cases the economic benefits of the antidumping duty are purely coincidental because a duty would be useful regardless of the pricing practices of the targeted exporters. See Dick (1991). It is fair to say that an academic consensus now exists to the effect that antidumping law is welfare-reducing in general. Some empirical work has been done on the welfare costs of antidumping measures as well, such as that in Messerlin (1989) and the papers collected in Lawrence (1998).

Although the normative analysis of antidumping law is well developed, rather little has been done on the positive side. If antidumping policy is so foolish, why is it also so durable, particularly in the WTO environment where nations have mutually agreed to forego economically foolish policies on some other fronts? I am aware of no systematic work on this important topic.

Two other strands of literature deserve mention. One seeks to examine whether decisions in antidumping cases are made "on the merits," or whether political factors (such as measures of the political influence of petitioning industries) significantly influence outcomes. Finger, Hall, and Nelson (1982) initiated this literature with respect to antidumping cases in the United States. They find that political factors have significant influence in settings where agencies have considerable discretion, such as in the injury investigations by the U.S. International Trade Commission (ITC). Anderson (1993) questions this conclusion, finding that the ITC follows the statutory injury factors rather faithfully without regard to politics.

Finally, a number of writers have criticized the way that the ITC operationalizes the injury test under antidumping law. They note that Commissioners may confuse correlation with causation, and may rest their decisions on economic fallacies (such as the notion that underpricing by imports proves injury due to dumping when it in fact may simply reflect quality differences). They urge greater reliance on economic modeling to quantify the impact of dumping, including various simulation techniques. Representative papers in this genre include Knoll (1989) and Boltuck (1991).

5.3.2. Trade-related aspects of intellectual property

The central issues of intellectual property law and policy are addressed elsewhere in this Handbook. The focus here is on the WTO Agreement on Trade Related Aspects of Intellectual Property (TRIPs), negotiated during the Uruguay Round of the mid-1990's. TRIPs requires all WTO members to afford a range of intellectual property rights on a non-discriminatory basis to domestic and foreign nationals, including patent, copyright and trademark protection. Subject to a few exceptions and to some transition provisions for developing countries, it harmonizes national intellectual property law to a considerable degree on issues such as what is patentable or copyrightable, and the duration of patent protection.

Much of the early writing on the TRIPs proposal focused on the welfare economics of globalizing intellectual property rights, particularly in the patent arena. The literature posed two central questions: does the globalization of intellectual property protection enhance global welfare, and does it enhance the welfare of developing countries? The answers depend importantly on what one assumes to be the set of available policy instruments for inducing the production of intellectual property. It is well known, for example, that patent protection involves an economic tradeoff between the costs of (transitory) monopolization of patented products, and the incentives to innovate. In theory, public subsidies to innovation might achieve an optimal rate of technical progress without the monopoly distortion of patent protection, although in practice the subsidization of invention is quite limited, perhaps for good reasons related to the capacity of governments to direct subsidies appropriately or to the costs of such a process. Deardorff (1992) assumes that public subsidies are infeasible, and analyzes the welfare economics of patent protection in a simple two-country model. It reveals the usual tradeoff between the monopoly distortion and the amount of innovation, here as a function of the geographic scope of patents. If one assumes that most of the innovation takes place in one country in the model (tracking the conventional wisdom about the North–South divide), his model further suggests that enlarging the scope of patent protection will tend to benefit the innovative country, and will tend to reduce welfare in the other country, with the effect on world welfare ambiguous in the general case.

The issue is not quite so simple, of course, because intellectual property protection can affect welfare in other ways besides simply the monopoly/innovation. It is conceivable, for example, that greater patent protection facilitates technology transfer from North to South, which in turn may facilitate the growth of industries in the South with positive spillovers. The ultimate question as to how an agreement like TRIPs affects economic development is a complicated empirical one, therefore, raising issues beyond the scope of this short survey. A thoughtful assessment of the issues and evidence is that of Maskus (2000).

The notion that TRIPs is likely injurious to developing countries, however, has been widely credited. A number of economic commentators suggest that TRIPs works an unwarranted transfer of rents from South to North, and should never have been bundled with market access agreements in the trade area, which standing alone are a source of mutual welfare gains. See, for example, Bhagwati (2002).

Why did the WTO membership nevertheless agree to TRIPs? Scotchmer (2004) offers a political economy model of intellectual property treaties. She posits two countries, each maximizing national economic welfare. They must decide on the scope of patent protection for inventions, on whether to produce innovations in the public sector (at a cost premium) in lieu of offering patent protection, and on whether to afford any patent protection to the products of the other country ("national treatment" in WTO parlance). Her model identifies a number of externalities that may (or may not) be successfully addressed by treaties. For example, if a nation unilaterally grants patent protection to inventions created abroad, it experiences an outflow of rents on all such inventions that would have been invented anyway. Only if the benefits from the induced increase in innovation abroad are sufficient to offset that loss will nations unilaterally extend protection, and hence they may fail to do so when it is globally efficient. If this situation arises, the two nations may then improve their joint welfare by an agreement for the reciprocal extension of intellectual property protection. Similar externalities arise in the decision to afford public financing of innovation, since the nation that finances innovation bears the costs but the benefits are enjoyed partly by foreign consumers. Scotchmer explores various cases involving countries that are symmetrical or asymmetrical in their capacity to innovate and in their consumer population. The externalities vary in their details by case, but in all cases some opportunity arises for the countries to improve on Nash strategies through international agreement.

Grossman and Lai (2004) make a number of related points in a model that presupposes a binding national treatment obligation on all nations. They utilize a two-country model capturing stylized assumptions about the North–South divide on intellectual property rights. In Nash equilibrium, countries fail to account for the benefits that greater intellectual property protection will yield to foreign consumers, and the level of intellectual property protection chosen will tend to be inefficiently low. International agreements can enhance welfare by strengthening intellectual property rights, although there is no need for "harmonization" to occur as it has to a large extent within the WTO system. Efficiency can be achieved through a wide range of policies trading off greater intellectual property protection in one nation for lesser degrees of protection in the other.

Scotchmer's discussion makes the further point that not all efficient international agreements will be reached if intellectual property policy is the only issue on the table. Globally efficient agreements may not lie in the bargaining core when the costs and benefits of agreements are asymmetric. This observation offers a key insight into the origin of TRIPs, which was accepted by all WTO member states, including less innovative states that are unlikely to be net beneficiaries of TRIPs standing alone. The inclusion of TRIPs within WTO auspices is a clear illustration of the issue linkage phenomenon discussed at some length earlier. Developing countries, which perceive themselves in large measure as consumers rather than producers of intellectual property, were opposed to the idea of TRIPs during much of the Uruguay Round negotiations. Intellectual property interest groups in the developed countries, representing such industries as pharmaceuticals, sound recording and filmmaking, were the principal proponents of TRIPs. The ultimate willingness of developing nations to accept TRIPs was a consequence in large part of their ability to obtain market access concessions on other sectoral issues, such as textiles and agriculture. Their acquiescence suggests that as a whole, developing countries believed that they gained more from accepting TRIPs and its quid pro quo than by rejecting it. This observation calls into question some of the criticism of TRIPs from the development perspective—its impact cannot be assessed in isolation, but must be considered in relation to the trade concessions that developing countries received in exchange for it.

Since the entry into force of the TRIPs agreement, perhaps the most controversial issue relating to implementation has involved its effect on access to "essential medicines" in developing countries, particularly drugs for the treatment of AIDS in the pandemic regions of Africa. Political controversy over the matter led the WTO membership to adopt a declaration during its Doha ministerial meeting, underscoring the right of member states to engage in compulsory licensing of patents for domestic production and consumption, and to determine their own policies regarding the "exhaustion" of intellectual property rights. The latter issue concerns the ability of patentholders to prevent the entry of so-called "parallel imports" into a market where they hold a valid patent. If they cannot, goods sold by a patentholder at a low price in one market can be exported or re-exported to another market where the patentholder sells at a higher price, thus undercutting efforts at price discrimination.

Much has been written about the economic policy issues associated with the Doha declaration, largely revolving around two central observations. First, any weakening of intellectual property rights may yield benefits from lower cost access to existing pharmaceuticals in poorer countries, but will come at some price to the incentives for pharmaceutical research, especially as to tropical diseases such as malaria and sleeping sickness that are largely confined to developing countries. Second, pharmaceutical companies are well aware that demand conditions vary across national markets, and are actively engaged in price discrimination to exploit it. The global welfare effects of such price discrimination are perhaps ambiguous for the usual reasons, but it is clear that policies undermining price discrimination (such as a permissive policy toward parallel imports) will tend to *raise* prices in the markets with weaker and more elastic demand, most likely the poorer nations. Rather than weakening intellectual property rights in developing nations, a useful response to the current problems may be to *encourage* international price discrimination (by disallowing parallel imports). Such a policy may provide developing countries with more affordable access to drugs that are marketed globally (such as AIDS medications), while preserving whatever incentive may exist for the development of new drugs to address tropical diseases. A related proposal, put forward by Jean Lanjouw, would allow pharmaceutical companies to obtain patents on new drugs in either developed or developing countries, but not both (while prohibiting imports of generic drugs from abroad during the life of a patent). Lanjouw's proposal has the advantage of ensuring that drugs are available at marginal cost in developing countries if they are patented in developed countries (the likely choice for new AIDS drugs, for example), while leaving open the option for pharmaceutical companies to obtain patents in the developing world for new drugs addressing tropical diseases. For a flavor of the debate and the literature, see Scherer and Watal (2002) and Sykes (2002).

6. International investment law

Just as national policies toward trade in goods and services cause externalities, so too can national policies toward factor flows. As in the trade area, the policy choices of a "small' country will not affect factor suppliers or purchasers abroad, where "smallness" simply means that the nation has no power to affect factor prices or returns on world markets. But in many practical settings, national policies to restrict factor flows can and do affect prices. If a "large" country imposes a tax on foreign direct investment within its territory, for example, it can extract some of the rents on inframarginal investments.

Interestingly, multilateral agreements regarding factor flows are of minimal significance. Immigration policy has long been viewed as the province of national governments, and has been little affected by multilateral agreements (save for limited commitments within GATS to permit the temporary movement of persons to establish service providers). The WTO has the Agreement on Trade-Related Investment Measures (TRIMs), but that agreement is no more than a restatement of GATT rules regarding trade in goods implied by GATT Article III (national treatment) and GATT Article XI (elimination of quantitative restrictions).[8] The OECD put forth a proposal in the 1990's for a multilateral agreement on investment, but it failed in the face of anti-globalization activism.

[8] For example, TRIMs prohibits restrictions on foreign investors that require their operations to buy domestic rather than imported input products. Such restrictions represent a condition affecting internal sale that favors domestic over imported goods, and violates GATT Article III(4).

Investment issues have been the subject of numerous bilateral agreements, however, dating back to early Friendship, Commerce and Navigation Treaties. Recent years have seen a rapid proliferation of so-called Bilateral Investment Treaties (BITs), and there are now well over a thousand of them. BITs are generally negotiated between a developed country and a developing country, and their primary purpose is the protection of foreign investors against expropriation by the government. They provide investors with a right to prompt compensation in the event of expropriation, and also typically provide that disputes may be referred to neutral arbitration. BITs are something of an anomaly in international law in that they give private actors the right to bring an action against a government alleged to have violated the treaty (i.e., private "standing"), and further require a government adjudged to have breached its obligations to pay money damages to the aggrieved investor. Investor rights provisions along these lines are also to be found in NAFTA. Many BITs also require national treatment toward investors of the other party, thus precluding various types of discriminatory measures that do not violate WTO rules.

Guzman (1998a) reviews the history of BITs, and focuses on a historical puzzle regarding their rising popularity. He notes that the protection of investor rights in developing nations has waxed and waned over the past century. After the emergence of customary international law requiring prompt and adequate compensation for any expropriation of the property of an alien, a movement within the United Nations led to the passage of a General Assembly resolution that weakened the right to compensation and restored national "sovereignty" over domestic resources. Developing nations overwhelmingly supported this resolution. Not long afterward, however, BITs became popular and many developing countries flocked to them, agreeing under BITs to waive their "sovereignty."

Guzman argues that the apparent contradiction in the behavior of developing countries may be explained by the fact that they were acting collectively in the United Nations, but were forced into unilateral action by the push for BITs. He suggests that the opportunity to selectively expropriate investors amounts to a tax on investment that benefits developing countries because of their collective market power over investment opportunities. Accordingly, they sought to strengthen that power when given the opportunity to do so collectively within the United Nations. But whey they were approached on a bilateral basis and asked to sign a BIT, they faced a Prisoner's Dilemma. Each nation could attract more investment by committing not to expropriate investors, but the private gain from such a commitment was largely a transfer from other developing nations. The growth of BITs thus represents a cascade of defection from the collective interest of developing nations. Guzman thus concludes that although BITs may be globally welfare enhancing because they represent a retreat from the exercise of market power by developing nations, they may well have lowered the welfare of the developing world. An alternative hypothesis, of course, is that developing countries came to realize that their strategy in the United nations was a mistake, and that retaining an opportunity to expropriate raised their cost of capital by more than the gains from any expected expropriation (perhaps because of the risk aversion of investors). Although Guzman's thesis is surely an intriguing one, it likely does not represent the final word on the matter.

Elkins, Guzman, and Simmons (2004) examine the diffusion of BITs empirically. They employ three measures to capture the degree to which developing countries compete with each other for capital—the overlap in the countries' export markets, the overlap in their export sectors, and the similarity of their labor force and infrastructure. With all three measures, they find that the likelihood of a nation entering a BIT is greater if a "competitive" nation has already entered one. They take their results as support for the proposition that BITs arise as a commitment or signaling device by developing nations seeking to lower the cost of foreign capital.

Sykes (2005b) examines the related question of why, in dramatic contrast to trade agreements, international investment agreements routinely afford investors a private right of action for money damages in the event of a breach of the agreement. He suggests that investment agreements are motivated by a desire for commitments (or signals) from developing country governments to foreign firms, ensuring protection for the sunk investments of those firms. To reduce the risk premium on imported capital as much as possible, developing country governments must credibly promise monetary compensation for expropriation to the investors themselves. Investors cannot count on their home governments to protect them as well in the absence of a private right of action because, among other things, some investors may be politically ineffective. Host countries are thus better off by making a credible promise of compensation to all investors, whether or not they are well-organized politically in their home countries. In the case of trade agreements, by contrast, officials in importing nations do not perceive any direct benefit from reducing the risks to foreign exporters. Increased imports are viewed as an "evil," and are tolerated only because they are accompanied by market access concessions abroad that well-organized export groups demand. Parties to trade agreements thus have little interest in improving export opportunities for exporters that are not well-organized. For this reason, private enforcement actions may reduce the political utility of trade agreements because the beneficiary of the enforcement action may be a poorly-organized export group, and the action may burden a well-organized import-competing industry. Governments can avoid the loss of political welfare that can arise in such cases by reserving to themselves the right to decide whether a case should be brought.

Been and Beauvais (2003) address another important issue that has arisen with respect to investor rights treaties—the definition of "expropriation." Recent arbitral decisions pursuant to the NAFTA investor rights provisions have suggested that "expropriation" under NAFTA rules may include what U.S. legal scholars sometimes refer to as regulatory takings. In one case, for example, an American waste disposal company won a multimillion dollar award against Mexico after operating permits for a disposal facility that it was constructing in Mexico were denied on environmental grounds. The arbitrators concluded that the Mexican government had misled the investor and that it had reasonably relied on assurances that the facility would be allowed to operate.

Been and Beauvais criticize this and other decisions in so far as they represent a trend toward incorporating regulatory takings into the protection against expropriation. They review the literature on takings generally, which contains two economic arguments for compensating those injured by a regulatory taking. One strand, often associated with Epstein (1985), argues that compensation is necessary so that governments will internalize the costs of their regulatory decisions and make them efficiently. Been and Beauvais echo some familiar criticisms of this argument, making the central point that the argument for cost internalization by government agencies is weakened by the fact that those agencies do not reap the benefits of their regulatory actions. To require them to bear the costs may then distort rather than correct incentives. Another strand of literature, often associated with Blume and Rubinfeld (1984), argues for compensation as an insurance mechanism. Been and Beauvais suggest that public compensation for regulatory takings is difficult to justify on this basis given the moral hazard problem that it creates. Economically justifiable insurance, they suggest, will generally be supplied by the market and should be procured there (they note the existence of market-based political risk insurance as an example). Finally, they question any suggestion that regulatory takings compensation will benefit nations by lowering their cost of capital enough to exceed the costs of compensation—were that so, nations would regularly offer clear assurances of compensation for regulatory takings and we simply do not observe them. Accordingly, they conclude that the broadening of the concept of "expropriation" to encompass regulatory takings is unwise within NAFTA and in other international settings (such as BITs).

7. International antitrust

The last decade has witnessed an explosion of commentary on international antitrust issues, spurred by an initiative on the part of some WTO member states to bring competition policy under the WTO umbrella. Much ink has been spilled over the wisdom of a global competition policy agreement, and over the question whether the WTO is an appropriate place to locate it. To date, however, only limited international cooperation on antitrust issues has emerged, mainly through bilateral agreements on the exchange of information.

Proponents of an international competition policy agreement often make their case by pointing to policy externalities, the existence of which has long been recognized. See Ordover and Sykes (1988). For example, a single price monopolist is a source of deadweight loss in a closed economy. But if the monopolist is an exporter, at least some of its monopoly profit represents a transfer from foreign consumers. From the national welfare perspective of the exporting country, therefore, the existence of the monopolist may be welfare enhancing even though it assuredly reduces global welfare in the aggregate. If the antitrust enforcement authorities in the exporting country give more weight to domestic welfare than to global welfare, a seemingly plausible scenario, they may then decline to act against the monopolist even if they might do so in a closed economy situation. The same may be said about the decision to act against a domestic cartel, or against domestic members of an international cartel. Anecdotal evidence suggests that governments indeed seem to favor national welfare over global welfare at times—the Webb-Pomerene Act in the United States, for instance, exempts export cartels from antitrust prosecution. If the nations that are injured by resulting anticompetitive practices are for some reason unable to do anything about it, perhaps because of difficulties in securing jurisdiction or in mounting a credible threat to sanction the firms in question, behavior that reduces global welfare may persist.

The reverse problem can also arise. Imagine a merger between two firms in a concentrated industry, and assume that the merger increases global welfare because it produces efficiencies that exceed any additional deadweight costs due to the higher oligopoly margins that may result from increased industry concentration. From the perspective of a nation that imports the goods produced by that industry, however, the merger may appear undesirable because the importing nation sees only the possibility of higher prices, while the efficiency gains may be captured primarily by the foreign shareholders of the merging firms. If the importing nation has the ability to block the merger, therefore, it may choose to do so even though the merger enhances global welfare. The general lesson is that the effects of imperfect competition on the welfare of individual nations turns critically on the national identity of the consumers and shareholders in an industry and on the division of surplus among them. It follows immediately that antitrust policy, which can alter conditions of competition in imperfectly competitive industries, has potentially important external effects in industries with significant international trade.

Commentators such as Fox (1999) and Guzman (1998b) proceed from such observations to argue that some degree of international cooperation on antitrust is desirable. Even accepting this premise, however, a possible difficulty emphasized by Guzman is the problem of an empty core. He doubts that all nations would gain from a competition policy agreement that imposed sound principles of antitrust policy. He thus argues for issue linkage as a possible solution, and thus for embedding competition policy within the broader WTO umbrella. Opponents of extensive WTO involvement, including Fox, express concern about the ability of its dispute resolution mechanism to address antitrust issues adequately, and fear that the WTO tradition of accommodating protectionist pressures may undermine the pursuit of efficient antitrust policy. Still other commentators, such as Wood (1999), oppose any form of multilateral antitrust agreement. They reason that no consensus exists on the proper principles of antitrust law. If Chicago school thinking has come to pervade American policy, it has by no means swept the debate in venues such as Europe and Canada. The compromises necessary to reach an international agreement on key principles, therefore, might undermine the core values of national antitrust enforcers.

The evident lack of consensus on many antitrust principles has led some observers to propose more modest agreements, perhaps imposing a basic national treatment obligation (an obligation to treat foreign firms no less favorably than domestic firms), and an agreement to prohibit certain hard core cartel practices. But even these seemingly narrow commitments are not without their problems. Antitrust enforcement often involves difficult and rather subjective judgments—will a merger lead to substantial efficiencies, and will the increase in concentration cause prices to rise importantly? If national authorities were accused of shading these judgments to favor their domestic firms over foreign interests in violation of a national treatment obligation, it is not at all clear how a dispute process could confidently determine the truth of the allegation. Likewise, although a consensus may exist that hard core cartel practices such as price fixing are undesirable, disagreement may well arise over the question of when price fixing is present. The long-standing controversy in the United States over the line between price fixing and mere "conscious parallelism," and the "plus factors" that are required to prove the former, illustrates the problem. At least in part because of such difficulties, the prospects for substantially greater international cooperation on antitrust issues appear rather bleak at this writing.

8. Human rights law

Despite its importance within the legal academy, virtually nothing has been written from a theoretical perspective by economically oriented scholars on international human rights law. Only Goldsmith and Posner (2005) address it briefly.

The subject matter of human rights law is primarily the treatment of domestic nationals by their own governments. Both treaties and customary rules create a number of "rights" in this regard, ranging from prohibitions on genocide and torture to rules prohibiting discrimination on the basis of race or gender. It is not immediately obvious why such matters, which are seemingly internal to each state, should become a subject of international law at all. The best theory of an "externality" is perhaps altruism—citizens of foreign states care about the welfare of oppressed people abroad to a degree, and are willing to expend resources to help them. Certainly the recent history of humanitarian interventions in various settings offers some support for the existence of this externality. Yet, unlike the externalities discussed in earlier sections of this chapter, there is little reason to think of this externality as reciprocal: citizens of the United States may care about the fate of repressed minorities in Serbia, Iraq or Afghanistan, but there is little reason to think that citizens of those countries simultaneously worry about human rights violations in the United States. Thus, human rights agreements do not fit well within the standard model of international agreements under which each signatory gains from the elimination of a reciprocal externality, and it is something of a puzzle as to why nations with weak human rights regimes would sign them (absent side payments, which we do not seem to observe). Likewise, international human rights treaties generally lack a formal enforcement mechanism, and it is difficult to see how such agreements can be self-enforcing. Threats of mutual defection are useless if a violator state does not care about defection by another.

Hence, a positive theory of human rights agreements must look to other considerations. Some of the explanations that Goldsmith and Posner offer for customary international law may have purchase. Human rights agreements may simply memorialize a coincidence of interest among most signatories, stating principles to which their domestic legal systems already conform. Repressive states may sign such treaties as well thinking that violation is essentially costless, and that a public commitment to human rights may have some modest political benefit. Coercion is another possible explanation. Liberal states may create human rights treaties for the purpose of altering the global political dynamic when they contemplate the use of force or economic sanctions. Support for coercive actions against repressive states may be enhanced if the repressive states can be said to be in violation of international law. Related, human rights violations may become the predicate for war crimes trials after a successful military intervention, and may then have some value for the deterrence of rogue leaders. Finally, political scientists have noted that where human rights treaties are binding under domestic law, such as certain intra-European agreements, they may be used by the leaders of newly democratic states to tie the hands of future leaders and thereby to discourage a return to authoritarianism.

In light of these considerations, it is perhaps unsurprising that violations of human rights agreements are rampant, as Goldsmith and Posner document. The finding in the empirical political science literature that ratification of human rights agreements seems to have little effect on the behavior of ratifying states is also unsurprising. See Hathaway (2002). Human rights treaties in the main commit liberal states to behave as they would anyway. Whenever a treaty would require a liberal state to change its policies, they typically sign only with a reservation (such as certain reservations taken by the United States to preserve its right to use capital punishment). Rogue states and other repressive regimes may or may not sign human agreements, but when they do so they generally have no intention of altering their behavior to comply.

9. Conflicts of law

Conflicts of law, a traditional focus of academic work on "private international law," arise when aspects of a legal dispute may be governed by two or more different legal rules. The issue usually arises when a dispute involves parties who are citizens of different states. Although the subject is often discussed with reference to conflicts of law among domestic jurisdictions, it has an obvious international counterpart. Consider, for example, the events some years ago in Bhopal, India involving an explosion at a chemical plant owned by an American company (Union Carbide). Suppose that injured parties wish to file a lawsuit seeking compensation—should the suit be governed by Indian law (the law of the jurisdiction where the accident occurred), U.S. law (the law of the jurisdiction in which the defendant is incorporated), or the law of some other jurisdiction? The question is an important one because the choice of governing law may significantly affect the plaintiff's chances for victory, or the damages that the plaintiff can collect. It may also have other allocative consequences as discussed below.

To the extent that local laws are tailored to reflect local conditions (such as the residents' willingness to pay for safety), a simple argument may be made for applying the law of the jurisdiction where the harm arose. But such a rule may lead to strategic incentives that create important inter-jurisdictional externalities. The work that has been done on the subject in the domestic context focuses on the question whether states will manipulate their choice of law rules to extract surplus from other states, and whether that in turn will adversely affect other rules of substantive law. The two leading papers explore these questions with particular reference to state products liability law. McConnell (1988) considers the consequences of the traditional U.S. rule for choosing the law to govern an accident—the rule that the court should apply the law of the jurisdiction in which the accident occurred (the "territorial rule"). He suggests that under the territorial rule, states will have an incentive to adopt substantive rules of accident law that inefficiently extract surplus from other states: Injured parties will tend to sue in their home state, McConnell suggests, so that most suits will involve in-state plaintiffs, yet many defendants will be out of state companies. On average, therefore, an inefficiently pro-plaintiff law can transfer wealth into the state as long as sellers cannot adjust their prices to reflect liability costs in each state. Out of state product manufacturers cannot do so, argues McConnell, because arbitrage across jurisdictions prevents price discrimination. Thus, the equilibrium is one in which all states tend to employ suboptimally pro-plaintiff laws. They would benefit from cooperation to avoid the problem, but cooperation is difficult when the number of states is large.

Hay (1992) challenges this analysis, noting that it rests crucially on the assumption that states must employ the law of the jurisdiction where the accident occurs. An alternative choice of law rule has emerged in the United States under the rubric of "interest analysis," which Hay suggests is sufficiently malleable to allow states to choose between the law of the state in which the accident happened and the law of the state of the defendant. He argues that states will jettison the traditional rule, and use interest analysis to choose whichever law is more favorable to their in-state plaintiff. Their incentive to adopt pro-plaintiff rules of substantive law then diminishes because such rules become a double-edged sword. They benefit in-state plaintiffs in suits against out-of-state firms as before, but they can also be invoked by out-of-state plaintiffs in their home states to disadvantage in-state firms who do business there and become defendants. He further notes that once all states switch to interest analysis, states with relatively pro-plaintiff laws will cause their firms to relocate to jurisdictions with more pro-defendant rules. Hay does not fully model the resulting equilibrium, but suggests that this process will exert powerful discipline on the tendency of states to adopt unduly pro-plaintiff laws.

This work on domestic issues suggests some principles that bear on the international context as well. States can exploit out-of-state sellers of goods and services (or investors) through pro-plaintiff rules of substantive law, or pro-plaintiff conflicts of law rules, only if the out-of-state entities are unable to adjust their prices to recover differences in costs across jurisdictions *or* if they have made sunk investments in the state and the legal rule comes as a surprise. As to the first problem, arbitrage may be less of a constraint on international price discrimination given transportation costs and trade barriers, but the problem may still arise. As to the second, unanticipated changes in liability rules are certainly possible, although prices may adjust going forward to restore a competitive return. Investors fearful of unanticipated liability may also charge a risk premium in capital markets, creating an incentive for host states to signal their intention to maintain existing rules if they can do so credibly. It is questionable whether BITs have any force in this regard, as it is doubtful that a change in a liability rule would be deemed "expropriation."

Even if investors anticipate all liability rules, however, inappropriate choice of law rules can distort investment (as well as trade patterns). To return to the Bhopal example, suppose that U.S. law is more generous toward plaintiffs than Indian law. Suppose further that Indian law allows Indian plaintiffs to choose between Indian law and the law of the jurisdiction in which the defendant is incorporated in all suits against foreign companies doing business in India. U.S. firms will then be at a competitive disadvantage relative to Indian firms and perhaps third country firms as well. The result will be to encourage investment, production and/or imports from firms that may be less efficient than U.S. firms. Put differently, if the choice of law rules result in non-uniformity in the rules applied to companies from different countries, the problem of trade diversion arises along with the distortion of investment patterns. The same problem arises if Indian plaintiffs are allowed to come to U.S. courts and invoke U.S. law against U.S. companies when they cannot invoke the same rules against competitors of U.S. companies. These points may be found in Sykes (2006b), which demonstrates formally that discrimination in the substantive tort rules applicable to exporters or investors of different nationalities can lower global welfare, even if the alternative to discrimination is the uniform application of tort law that is inadequate to induce proper levels of precautions. The open question is whether the sorts of rules that can produce these distortions will emerge and persist in equilibrium, thus justifying some sort of international cooperation to address the problem.

The scope of the issues raised by choice of law problems is, of course, far broader. In many substantive areas of law, the decision by one jurisdiction to apply its law to transactions or events outside its jurisdiction will create important externalities. Economically-oriented scholars have begun to wrestle with some of these issues in a few areas, such as securities regulation. See Romano (1998). But much remains to be done. Guzman (2002b) provides an overview of the issues and some general principles for consideration, with particular attention to issues arising in antitrust, securities and bankruptcy.

10. The international commons: the example of fisheries

I conclude with a short note on an important source of international externalities not yet noted in this chapter: the problem of incomplete property rights. Many valuable resources are unowned, leaving no actor with an incentive to protect them from uneconomic deterioration or more generally to maximize their value. The problem arises with the atmosphere, the oceans, space, Antarctica, the common pasture, oil pools, and innumerable other resources. I will use fisheries to illustrate the problem because of the rather interesting (and also discouraging) body of international law on the matter.

To crystallize the nature of the externality problem, consider a two-period model of a fishery adapted from classic analyses by Gordon (1954) and Cooper (1975). For a single species fishery, let H_t denote the total number of hours of fishing labor required to catch Y_t fish in fishing season t, and let s_t denote the stock of fish at the beginning of season t. The relation between H, Y, and s is given by $H_t = f(s_t)(Y_t)^2$, where $f' < 0$ and $f'' > 0$. Thus, the amount of time per fish invested in fishing increases as the catch increases (reflecting the increasing scarcity of fish), and decreases as the initial stock of fish increases. The beginning of season stock is determined by the size of the stock at the end of the last season plus new "births" given by a transition function that relates the number of new fish added to the fishery between seasons to the end of period stock: $s_t = s_{t-1} - Y_{t-1} + B(s_{t-1} - Y_{t-1})$. We assume that $B' > 0$ at least up to the point of some ecological limit on the fishery, and $B'' < 0$. The price of fishing labor is unity, and the price of a fish, assumed constant over time for simplicity, is p. The social rate of discount is r, and let $\delta = 1/(1 + r)$.

Efficient use of this price-taking fishery requires that the discounted value of profits from the fishery be at a maximum. Using the relation between fishing hours and the catch, the profit function may be written as:

$$\pi = pY_1 - f(s_1)(Y_1)^2 + \delta[pY_2 - f(s_2)(Y_2)^2].$$

The optimization problem is then to select the size of the catch in each fishing season to maximize this function, subject to an initial condition on the stock of fish and to the transition function. The solution to this dynamic programming problem is found by backwards induction. The maximization of season two profits given the initial stock in season two simply requires that price be set equal to marginal cost in season two: $p = 2f(s_2)Y_2$, which allows season two profit (π_2) to be expressed as a function of s_2. Substituting into the profit function above and using the transition function and the initial condition on the stock, the first order condition for the optimal catch in season one (assuming a positive amount of fishing is optimal) may be written as:

$$p = 2f(s_1)Y_1 - \delta(\partial\pi_2/\partial Y_1).$$

This condition also states that price must equal marginal cost, but has two components—the season one marginal cost given the initial stock, plus the effect of marginal fishing in season one on discounted profit in season two (via its effect on the season two stock). Assuming that the fishery is below its ecological limit, the second term on the right hand side is positive. Denoting the first term on the right hand side as "short-term" marginal cost, it is clear that properly computed marginal cost exceeds short-term marginal cost, and that the optimal amount of fishing is lower than would be implied by an equality between price and short term marginal cost.

The relationship between this condition for efficient use of the fishery, and the market equilibrium that actually emerges, will depend on the market structure of the industry that uses the fishery. Imagine, for example, that the rights to use the fishery are owned by a single profit-maximizing company. Such a company would confront the profit maximization problem stated above, and would exploit the fishery efficiently (the company's "monopoly" over the fishery causes no inefficiency because it is a price taker in the market for fish).

As the number of companies using the fishery increases, however, externalities emerge within each season and over time. To isolate the first, assume that each company in each season chooses its fishing effort in Cournot–Nash fashion, taking the fishing effort of other companies as given. Each company will equate price to its private marginal cost, which will incorporate the marginal effect of its own fishing on the hours required to catch fish. But it will neglect the fact that its own fishing also increases the costs to other companies of catching fish. The size of this external effect will tend to be greater as the number of companies using the fishery increases. In the limit, very small companies may behave as if the effect of their own fishing on the hours required to catch fish is zero. The equilibrium for this limiting case may be derived readily (I suppress time subscripts for simplicity). Let y and h denote the catch and hours investment of a small company. Each small company confronts the profit function per season of $\pi = py - h$, where y is given by the average productivity of fishing (Y/H) times its hours of fishing, and average productivity is taken to be fixed by each company. Equilibrium requires zero profits, which in turn implies that $p = Y/H$. The last equation states that in equilibrium, price will equal the *average* cost of fish, in contrast to the condition for efficiency, which requires price equal to marginal cost. Because average cost lies below marginal cost, the equilibrium involves excessive fishing. More generally, it is not difficult to show that with symmetric Cournot–Nash firms, the efficiency of the market equilibrium in each season (taking the initial stock as given) declines steadily as the number of firms increases—the greater the number of firms, the larger the wedge between private short term marginal cost and social short term marginal cost.

The intertemporal externality arises for essentially the same reason. With more than one firm in the fishery, each firm will take account of its own effects on the future stock of fish and its future costs of fishing, but will neglect the fact that greater fishing effort in the current season also raises costs for other firms in future seasons. Thus, focusing on the second term on the right hand side of the efficiency condition for season one, the private cost of fishing in period one via its effect on future profit is again less than the social cost. This further exacerbates the over fishing problem, and will tend to become more acute as the number of firms rises. In the limiting case, small firms may completely ignore the effect of their fishing effort on the future stock, and the market equilibrium in each season will involve price equal to short term average cost. Any number of trajectories are possible depending on the biology of the fishery, of course, but in the worst case scenario, the stock of fish may shrink steadily until the use of the fishery is no longer economical at all.

Possible corrective measures include all of the usual suspects. Restricting access to the fishery can improve the efficiency with which it is managed. Command and control regulation over the size of the catch can also help, as can taxes on fishing. The choice among these mechanisms turns on familiar considerations that I will not detail here.

To return to the subject of international law, however, the challenges of fisheries management can become all the more acute when no one regulatory authority has jurisdiction, so that a need arises to coordinate across jurisdictions. The history of international cooperation on fisheries management, however, is by and large unimpressive. The Law of the Sea grants each state dominion over a 12-mile area of territorial waters along its coastline, from which it has the right to exclude others. But very little has been done to coordinate fishing in the open oceans. The only substantial cooperation seems to arise when fishing directly or indirectly affects endangered species such as sea turtles or certain whales.

A number of factors may explain the lack of progress in this area. First, considerations of political economy suggest that support for fisheries management policies may be limited. Even if fishery restrictions would raise the discounted value of profits, existing fishing companies may not favor them. Their time horizons may be limited by the life of their sunk capital, and they may see the benefits of fisheries management inuring to their successors. Other beneficiaries include the future consumers of fish, who seem unlikely to exert much political pull toward cooperation. Companies such as canneries and other fish packing operations may weigh in favor of fisheries management, but the balance of political forces is unclear both within and across jurisdictions.

Second, the fisheries management problem can be highly complex. Interested parties may disagree on the optimal policy. Further, the policy in place in one fishery may have spillover effects on another—a reduction in over fishing in one area may raise prices or divert sunk capital resources to cause greater over fishing in another.

Finally, the enforcement of cooperation is likely difficult, particularly on the open oceans. Defections may be difficult to detect, and enforcement may be plagued by a free rider problem. The suspicion that another party is cheating may lead any agreement to unravel.

In light of these considerations, it is perhaps no surprise that the limited successes in the area have involved measures to protect endangered species. Such measures attract the support of the environmental lobby, which is the most striking example of diffuse interests coming together to organize politically. They may also attract private third-party enforcers such as Greenpeace, which has played a prominent role in publicizing violations of the Whaling Convention.

11. Conclusion

Economic analysis of international law is still in its infancy. With the exception of international trade, where economists with an interest in legal institutions have made considerable progress in understanding some of those institutions, formal analytic work is scarce and careful empirical work is rare. Few fields offer more opportunities for fruitful research.

Indeed, international law encompasses many subjects that I do not mention at all in this chapter, primarily because the economically-oriented work done on them to date is too sparse or non-existent. Consider such topics as the law of the sea, the law of aviation and outer space, and the law of arms control—each raises its own set of externality issues with an array of legal responses. Little or nothing has been written about them by economists. Over-arching arrangements such as the Vienna Convention on the Law of Treaties,[9] and the draft articles on the Responsibility of States for Internationally Wrongful Acts[10] have also attracted little attention, as have many aspects of domestic law that bear on international relations such as foreign sovereign immunity principles. Of course, many of the subjects that I do discuss above have vast numbers of open questions. Even in the trade area, which has received more scholarly attention than other subjects by orders of magnitude, many important legal subjects remain largely unaddressed (such as the reasons for the persistence of antidumping law, noted earlier, or the elaborate WTO rules for services and agriculture trade). The general theoretical approach considered in this chapter—identify the pertinent externalities or market failures and then consider the problem of how to design an optimal contract to address them—can bear fruit in most all of these subject areas.

Acknowledgements

I have received thoughtful comments on an earlier draft from Kyle Bagwell, A. Mitchell Polinsky, Eric Posner, Steven Shavell, and workshop participants in the Handbook of Law and Economics conference at Stanford Law School.

References

Anderson, K.B. (1993). “Agency Discretion or Statutory Direction: Decision Making at the U.S. International Trade Commission”. Journal of Law and Economics 36, 915–935.

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

Bagwell, K., Staiger, R.W. (1990). “A Theory of Managed Trade”. American Economic Review 80, 779–795.

Bagwell, K., Staiger, R.W. (1999). “An Economic Theory of GATT”. American Economic Review 89, 215–248.

Bagwell, K., Staiger, R.W. (2002). The Economics of the World Trading System. MIT Press, Cambridge, MA.

Bagwell, K., Staiger, R.W. (2004). “Multilateral Trade Negotiations, Bilateral Opportunism and the Rules of GATT/WTO”. Journal of International Economics 63, 1–29.

Bagwell, K., Staiger, R.W. (2006). “Will International Rules on Subsidies Disrupt the World Trading System?”. American Economic Review 96, 877–895.

Bagwell, K., Mavroidis, P.C., Staiger, R.W. (2002). “It’s a Question of Market Access”. American Journal of International Law 96, 56–76.

[9] Elements of the Vienna Convention have been addressed in some recent work by Fon and Parisi. See, e.g., Fon and Parisi (2006).

[10] Posner and Sykes (2007) offer a first cut at some of the issues raised by the draft articles on state responsibility, applying lessons from the economic literature on vicarious liability to the question of when states incur “state responsibility” under international law, and the related question of when individual criminal responsibility may be a useful supplement to state responsibility.

Bagwell, K., Mavroidis, P.C., Staiger, R.W. (2003). "The Case for Auctioning Countermeasures in the WTO". Mimeo.

Bagwell, K., Sykes, A.O. (2004). "Chile—Price Band System and Safeguard Measures Relating to Certain Agricultural Products". Mimeo, prepared for American Law Institute Project on Principles of World Trade Law.

Baird, D.G., Gertner, R.H., Picker, R.C. (1994). Game Theory and the Law. Harvard Press, Cambridge, MA.

Battigalli, P., Maggi, G. (2003). "International Agreements on Product Standards: An Incomplete Contracting Theory". NBER Working Paper No. 9533.

Bayard, T.O., Elliott, K.A. (1994). Reciprocity and Retaliation in U.S. Trade Policy. Institute for International Economics, Washington.

Been, V., Beauvais, J.C. (2003). "The Global Fifth Amendment? NAFTA's Investment Protections and the Misguided Quest for an International 'Regulatory Takings' Doctrine". New York University Law Review 78, 30–143.

Bhagwati, J. (2002). "Afterword: The Question of Linkage". American Journal of Interantional Law 96, 126–134.

Bhagwati, J.N., Hudec, R.E. (1996). Fair Trade and Harmonization: Prerequisites for Free Trade? MIT Press, Cambridge, MA.

Bhagwati, J.N., Panagariya, A. (1996). "Preferential Trading Areas and Multilateralism—Strangers, Friends or Foes?". In: Bhagwati, J., Panagariya, A. (Eds.), The Economics of Preferential Trade Agreements. American Enterprise Institute, Washington.

Blume, L., Rubinfeld, D.L. (1984). "Compensation for Takings: An Economic Analysis". California Law Review 72, 569–628.

Boltuck, R.D. (1991). "Assessing the Effects on the Domestic Industry of Price Dumping". In: Tharakan, P.K.M. (Ed.), Policy Implications of Antidumping Measures. North Holland, Amsterdam.

Brander, J.A. (1995). "Strategic Trade Policy". In: Grossman, G.M., Rogoff, K. (Eds.), Handbook of International Economics, vol. 3. Elsevier Science, Amsterdam and New York.

Carlsnaes, W., Risse, T., Simmons, B.A. (2002). Handbook of International Relations. Sage Publications, London.

Chang, H.F. (1995). "An Economic Analysis of Trade Measures to Protect the Global Environment". Georgetown Law Journal 83, 2131–2213.

Cooper, R. (1975). "An Economist's View of the Oceans". Journal of World Trade Law 9, 357–377.

Dam, K.W. (1970). The GATT: Law and International Economic Organization. University of Chicago Press, Chicago, IL.

Deardorff, A.V. (1992). "Welfare Effects of Global Patent Protection". Economica 59, 35–51.

Devereux, M.B. (1997). "Growth, Specialization and Trade Liberalization". Journal of International Economics 38, 565–585.

Dick, A. (1991). "Learning by Doing and Dumping in the Semiconductor Industry". Journal of Law and Economics 34, 133–159.

Dunoff, J.L., Trachtman, J.P. (1999). "Economic Analysis of International Law". Yale Journal of International Law 24, 1–59.

Eaton, J., Engers, M. (1992). "Sanctions". Journal of Political Economy 100, 899–928.

Eaton, J., Sykes, A.O. (1998). "International Sanctions". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 2. MacMillan, London.

Elkins, Z., Guzman, A.T., Simmons, B. (2004). "The Diffusion of Bilateral Investment Treaties, 1960–2000". Mimeo.

Epstein, R. (1985). Takings: Private Property and the Power of Eminent Domain. Harvard Press, Cambridge, MA.

Ethier, W. (2000). "Political Externalities, Nondiscrimination and a Multilateral World". Review of International Economics 12, 303–320.

Ethier, W. (2004). "Trade Agreements Based on Political Externalities: An Exploration, Third Version". Mimeo.

Finger, J.M., Hall, H.K., Nelson, D.R. (1982). "The Political Economy of Administered Protection". American Economic Review 72, 452–466.

Fon, V., Parisi, F. (2003). "Stability and Change in International Customary Law". Mimeo.

Fon, V., Parisi, F. (2006). "The Economics of Treaty Ratification". Mimeo.

Fox, E.M. (1999). "Competition Policy and the Millennium Round". Journal of International Economic law 2, 665–680.

Fudenberg, D., Tirole, J. (1991). Game Theory. MIT Press, Cambridge, MA.

Goldsmith, J.L., Posner, E.A. (1999). "A Theory of Customary International Law". University of Chicago Law Review 66, 1133–1177.

Goldsmith, J.L., Posner, E.A. (2005). The Limits of International Law. Oxford University Press, Oxford.

Gordon, H.S. (1954). "The Economic Theory of a Common Property Resource: The Fishery". Journal of Political Economy 62, 124–142.

Grossman, G.M., Helpman, E. (1995a). "Trade Wars and Trade Talks". Journal of Political Economy 103, 675–708.

Grossman, G.M., Helpman, E. (1995b). "The Politics of Free Trade Agreements". American Economic Review 85, 667–690.

Grossman, G.M., Helpman, E. (2002). Interest Groups and Trade Policy. Princeton University Press, Princeton, NJ.

Grossman, G.M., Lai, E.L.-C. (2004). "International Protection of Intellectual Property". American Economic Review 94, 1635–1653.

Grossman, G.M., Mavroidis, P.C. (2004). "United States—Definitive Safeguard Measures on Imports of Circular Welded Carbon Quality Line Pipe from Korea". Mimeo, prepared for American Law Institute Project on Principles of World Trade Law.

Grossman, G.M., Sykes, A.O. (2005). "A Preference for Development: The Law and Economics of GSP". World Trade Review 4, 41–67.

Guzman, A.T. (1998a). "Why LDCs Sign Treaties That Hurt Them: Explaining the Popularity of Bilateral Investment Treaties". Virginia Journal of International Law 38, 639–688.

Guzman, A.T. (1998b). "Is International Antitrust Possible?". New York University Law Review 73, 1501–1548.

Guzman, A.T. (2002a). "A Compliance-Based Theory of International Law". California Law Review 90, 1823–1887.

Guzman, A.T. (2002b). "Choice of Law: New Foundations". Georgetown Law Journal 90, 883–940.

Guzman, A.T. (2004). "The Design of International Agreements". Mimeo.

Hathaway, O. (2002). "Do Human Rights Treaties Make a Difference?". Yale Law Journal 111, 1935–2041.

Hay, B.L. (1992). "Conflicts of Law and State Competition in the Product Liability System". Georgetown Law Journal 80, 617–652.

Horn, H., Mavroidis, P.C. (2001). "Economic and Legal Aspects of the Most-Favored-Nation Clause". European Journal of Political Economy 17, 233–279.

Horstmann, I.J., Markusen, J.R., Robles, J. (2001). Multi-Issue Bargaining and Linked Agendas: Ricardo Revisited or No Pain No Gain, NBER Working Paper No. 8347, mimeo.

Hufbauer, G.C., Schott, J.J., Elliott, K.A. (1990). Economic Sanctions Reconsidered: History and Current Policy, 2nd edn. Institute for International Economics, Washington.

Jackson, J.H., Sykes, A.O. (Eds.) (1997). Implementing the Uruguay Round. Clarendon Press, Oxford, UK.

Johnson, H.G. (1953). "Optimum Tariffs and Retaliation". Review of Economic Studies 21, 142–153.

Kemp, M.C., Wan, H.Y. (1976). "An Elementary Proposition Concerning the Formation of Customs Unions". Journal of International Economics 6, 95–97.

Knoll, M.S. (1989). "An Economic Approach to the Determination of Injury Under United States Antidumping and Countervailing Duty Law". New York University Journal of International Law and Politics 22, 37–116.

Lawrence, R.Z. (1996). Regionalism, Multilateralism and Deeper Integration. Brookings Institution, Washington.

Lawrence, R.Z. (Ed.) (1998). Brookings Trade Forum. Brookings Institution, Washington.

Lawrence, R.Z. (2003). Crimes & Punishments?: Retaliation under the WTO. Institute for International Economics, Washington.

Leebron, D.W. (1997). "Implementation of the Uruguay Round Results in the United States". In: Jackson, J.H., Sykes, A.O. (Eds.), Implementing the Uruguay Round. Clarendon Press, Oxford.

Levy, P.I., Srinivasan, T.N. (1996). "Regionalism and the (Dis)advantage of Dispute Settlement Access". American Economic Review, Papers and Proceedings 86, 93–98.

Maggi, G. (1999). "The Role of Multilateral Institutions in International Trade Cooperation". American Economic Review 89, 190–214.

Maggi, G., Morelli, M. (2003). "Self-Enforcing Voting in International Organizations". NBER Working Paper no. 10102, forthcoming in American Economic Review.

Maggi, G., Rodriquez-Clare, A. (1998). "The Value of Trade Agreements in the Presence of Political Pressures". Journal of Political Economy 106, 574–601.

Maskus, K.E. (2000). Intellectual Property Rights in the Global Economy. Institute for International Economics, Washington.

Mavroidis, P.C. (2000). "Remedies in the WTO Legal System: Between a Rock and a Hard Place". European Journal of International Law 11, 763–813.

McConnell, M.W. (1988). "A Choice-of-Law Approach to Products-Liability Reform". In: Olson, W.K. (Ed.), New Directions in Liability Law. Academy of Political Science, New York.

Messerlin, P.A. (1989). "The EC Antidumping Regulations: A First Economic Appraisal, 1980–85". Weltwertschaftliches Archiv 125, 563–587.

Olson, M., Zeckhauser, R. (1966). "An Economic Theory of Alliances". Review of Economics and Statistics 48, 266–279.

Ordover, J.A., Sykes, A.O. (1988). "The Antitrust Guidelines for International Operations: An Economic Critique". In: Hawk, B. (Ed.), North American and Common Market Antitrust and Trade Laws. Matthew Bender, New York.

Posner, E., Sykes, A.O. (2005). "Optimal War and *Jus Ad Bellum*". Georgetown Law Journal 93, 993–1022.

Posner, E., Sykes, A.O. (2007). "The Economics of State and Individual Responsibility Under International Law". American Law and Economics Review; doi: 10.1093/ales/ahm001.

Rasmusen, E. (1989). Games and Information: An Introduction to Game Theory. Cambridge Press, Cambridge, UK.

Rodrik, D. (1995). "Political Economy of Trade Policy". In: Grossman, G.M., Rogoff, K. (Eds.), Handbook of International Economics, vol. 3. Elsevier Science, Amsterdam and New York.

Romano, R. (1998). "Empowering Investors: A Market Approach to Securities Regulation". Yale Law Journal 107, 2359–2430.

Rose, A.K. (2004). "Do We Really Know that the WTO Increases Trade?". American Economic Review 94, 98–114.

Sandler, T., Cauley, J. (1975). "On the Economic Theory of Alliances". Journal of Conflict Resolution 19, 251–276.

Sandler, T., Hartley, K. (2001). "Economics of Alliances: The Lessons for Collective action". Journal of Economic Literature 39, 869–896.

Scherer, F.M., Watal, J. (2002). "Post-TRIPs Options for Access to Patented Medicines in Developing Nations". Journal of International Economic Law 5, 913–939.

Schwartz, W.F., Harper, E.W. (1972). "The Regulation of Subsidies Affecting International Trade". Michigan Law Review 70, 831–858.

Schwartz, W.F., Sykes, A.O. (1996). "Toward a Positive Theory of the Most Favored Nation Obligation and its Exceptions in the WTO/GATT System". International Review of Law and Economics 16, 27–51.

Schwartz, W.F., Sykes, A.O. (2002). "The Economic Structure of Renegotiation and Dispute Resolution in the WTO/GATT System". Journal of Legal Studies 31, S179–S204.

Scotchmer, S. (2004). "The Political Economy of Intellectual Property Treaties". Journal of Law, Economics and Organization 20, 415–437.

Snidal, D. (1996). "International Political Economy Approaches to Institutions". International Review of Law and Economics 16, 121–137.

Snidal, D. (2002). "Rational Choice and International Relations". In: Carlsnaes, W., Risse, T., Simmons, B.A. (Eds.), Handbook of International Relations. Sage Publications, London.

Staiger, R.W. (1995a). "International Rules and Institutions for Cooperative Trade Policy". In: Grossman, G.M., Rogoff, K. (Eds.), Handbook of International Economics, vol. 3. Elsevier Science, Amsterdam and New York.

Staiger, R.W. (1995b). "A Theory of Gradual Trade Liberalization". In: Deardorff, A., Levinsohn, J., Stern, R. (Eds.), New Directions in Trade Theory. University of Michigan Press, Ann Arbor.

Sykes, A.O. (1989). "Countervailing Duty Law: An Economic Perspective". Columbia Law Review 89, 199–263.

Sykes, A.O. (1991). "Protectionism as a 'Safeguard': A Positive Analysis of the GATT 'Escape Clause' with Normative Speculations". University of Chicago Law Review 58, 255–305.

Sykes, A.O. (1992). "Constructive Unilateral Threats in International Commercial Relations: The Limited Case for Section 301". Law and Policy in International Business 23, 263–330.

Sykes, A.O. (1995). Product Standards for Internationally Integrated Goods Markets. Brookings Institution, Washington.

Sykes, A.O. (1998). "Antidumping and Antitrust: What Problems Does Each Address?". In: Lawrence, R.Z. (Ed.), Brookings Trade Forum, vol. 1998, pp. 1–43.

Sykes, A.O. (1999). "Regulatory Protectionism and the Law of International Trade". University of Chicago Law Review 66, 1–46.

Sykes, A.O. (2001). "'Efficient Protection' Through WTO Rulemaking". In: Porter, R.B., Sauve, P., Subramanian, A., Zampetti, A.B. (Eds.), Efficiency, Equity, Legitimacy: The Multilateral Trading System at the Millennium. Brookings Institution, Washington.

Sykes, A.O. (2002). "TRIPs, Pharmaceuticals, Developing Countries and the Doha 'Solution'". Chicago Journal of International Law 3, 47–68.

Sykes, A.O. (2005a). "The Economics of WTO Rules on Subsidies and Countervailing Measures". In: Appleton, A., Macrory, P., Plummer, M. (Eds.), The World Trade Organization: Legal, Economic and Political Analysis, vol. II. Springer Verlag, New York.

Sykes, A.O. (2005b). "Public vs. Private Enforcement of International Economic Law: Standing and Remedy". Journal of Legal Studies 34, 631–666.

Sykes, A.O. (2006a). The WTO Agreement on Safeguards. Oxford University Press, Oxford, UK.

Sykes, A.O. (2006b). "Transnational Tort Litigation as a Trade and Investment Issue". Mimeo.

Telser, L.G. (1980). "A Theory of Self-enforcing Agreements". Journal of Business 53, 27–44.

Trebilcock, M.J., Howse, R. (1995). The Regulation of International Trade. Routledge, London.

Trebilcock, M.J., Soloway, J. (2002). "International Trade Policy and Domestic Food Safety Regulation: The Case for Substantial Deference by the WTO Dispute Settlement Body under the SPS Agreement". In: Kennedy, D.L., Southwick, J. (Eds.), The Political Economy of International Trade Law: Essays in Honor of Robert E. Hudec. Cambridge Press, Cambridge, UK.

Viner, J. (1923). Dumping: A Problem in International Trade. University of Chicago Press, Chicago.

Viner, J. (1950). The Customs Union Issue. Carnegie Endowment for Peace, New York.

Wood, D.P. (1999). "International Law and federalism: What is the Reach of Regulation?". Harvard Journal of Law and Public Policy 23, 97–110.

AUTHOR INDEX OF VOLUME 1

n indicates citation in a footnote.

This page intentionally left blank

SUBJECT INDEX OF VOLUME 1