21st century REFERENCE SERIES

# 21st Century ECONOMICS

## A Reference Handbook

Edited by

Rhona C. Free

# 21st Century ECONOMICS

*A Reference Handbook*

# 21st Century ECONOMICS
## *A Reference Handbook*

Edited by
Rhona C. Free
*Eastern Connecticut State University*


Los Angeles | London | New Delhi
Singapore | Washington DC

*For information:*

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
33 Pekin Street #02-01
Far East Square
Singapore 048763

Printed in the United States of America

*Library of Congress Cataloging-in-Publication Data*

21st century economics: a reference handbook / editor, Rhona C. Free.
v. ; cm.
Includes bibliographical references and index.
ISBN 978-1-4129-6142-4 (cloth)
1. Economics—Handbooks, manuals, etc. I. Free, Rhona C. II. Title: Twenty-first century economics.
HB171.A19 2010
330—dc22 2010001776

This book is printed on acid-free paper.

10 11 12 13 14 10 9 8 7 6 5 4 3 2 1

| | |
|---|---|
| *Publisher:* | Rolf A. Janke |
| *Acquisitions Editor:* | Jim Brace-Thompson |
| *Developmental Editor:* | Sanford Robinson |
| *Reference Systems Manager:* | Leticia Gutierrez |
| *Reference Systems Coordinator:* | Laura Notton |
| *Production Editor:* | Kate Schroeder |
| *Copy Editors:* | Gillian Dickens, Matthew Sullivan, Karen Wolf |
| *Typesetter:* | C&M Digitals (P) Ltd. |
| *Proofreaders:* | Kristin Bergstad, Sally Jaskold |
| *Indexer:* | Joan Shapiro |
| *Cover Designer:* | Candice Harman |
| *Marketing Manager:* | Amberlyn McKay |

# CONTENTS

**PART VI. ECONOMIC ANALYSES OF ISSUES AND MARKETS**

# Preface

As I write this preface in mid-January 2009, interest in economics is at an all-time high. Among the challenges facing the nation is an economy with unemployment rates not experienced since the Great Depression, failures of major businesses and industries, and continued dependence on oil with its wildly fluctuating price. Americans are debating the proper role of the government in company bailouts, the effectiveness of tax cuts versus increased government spending to stimulate the economy, and potential effects of deflation. These are questions that economists have dealt with for generations but that have taken on new meaning and significance.

At the same time, economists' recent innovative approaches to analyzing issues and behavior and the expansion of economic analysis to nontraditional topics and arenas of activity have attracted new interest and attention from citizens and scholars. Economists are working with sociologists and psychologists in areas such as neuroeconomics, the economics of happiness, and experimental economics. They have applied economic analysis to sports, the arts, wildlife protection, and sexual orientation, in the process demonstrating the value of economic methods in understanding and predicting behavior in a wide range of human activities and in development of policies aimed at many social issues.

Economics is generally described as the study of resource allocation; or of production, distribution, and consumption of wealth; or of decision making—descriptions that sacrifice much for the sake of brevity. Within these relatively vague definitions lie fascinating questions and critical policy implications. Traditional economic analysis has been used to explain why people who are overweight tend to have lower incomes than those who are thin as well as why some nations grow faster than others. Economists have explored why people gamble even though they are likely to lose money as well as why stock markets respond in predictable or unpredictable ways to external events. They develop models to analyze how tax policies affect philanthropy and how managers of baseball teams can determine which players are worth their salary demands. The range of questions that falls within the domain of economic analysis is much broader (and more interesting) than those suggested by the traditional definition of the discipline.

The value of economic analysis in development of policies to address social issues is also much broader than generally perceived. Economists have played a critical role in the development of policies aimed at protecting endangered species and addressing global warming and climate change. They contribute to development of policies that will curb smoking, promote entrepreneurship, reduce crime, and promote educational quality and equality. And they also provide the theory and evidence that is applied in policy arenas more traditionally thought of as being in the purview of the discipline—managing unemployment, economic growth, and inflation; regulating industries to promote competition, innovation, and efficient outcomes; and developing tax policies and rates that achieve a range of possible objectives.

Encompassing analysis of traditional economic theory and topics as well as those that economists have only more recently addressed, this handbook will meet the needs of several types of readers. Undergraduate students preparing for exams will find summaries of theory and models in key areas of micro- and macroeconomics. Readers interested in learning about economic analysis of a topic or issue as well as students developing research papers will find introductions to relevant theory and empirical evidence. And economists seeking to learn about extensions of analysis into new areas or about new approaches will benefit from chapters that introduce cutting-edge topics.

Authors of chapters in this handbook come from universities and policy institutes around the world. They represent well-known, distinguished scholars as well as doctoral students working in areas that have only recently gained the attention of economists. Perhaps most important, they reflect the full range of ideological, methodological, and political orientations and perspectives present in the discipline. Authors were selected based on their ability to accurately, succinctly, and clearly address the chapter topic and to present a balanced approach to the analysis, but while some of the authors represent the orthodox approach to economics, others reflect one that is more radical. The willingness of economists from various schools of thought and with diverse orientations and perspectives to contribute chapters has certainly been of benefit to this book.

For most authors, the biggest challenge in writing their chapter will have been the need to avoid using calculus and to limit the level of technical and quantitative detail. Calculus is like shorthand for economists; it provides them

with a quick and efficient means of analyzing and explaining relationships between economic variables. It is the language economists are comfortable using; it is the language they were trained in. Furthermore, economists are accustomed to discussing empirical evidence in terms of results of complex statistical tests and econometric methods. But to make the book accessible to undergraduate students and to nonacademic readers, it was essential that models be presented only in graphical format, or with minimal calculus, and that empirical evidence be summarized in a way that does not require much background in statistics or econometrics. The authors have been remarkably successful in achieving this goal, which for most will have required very careful thought and many revisions. To help readers understand some of the empirical parts of the chapters, there is a separate chapter that reviews basic concepts in econometrics. Readers challenged by more difficult models and applications may find it helpful to read some of the earlier chapters that review basic theories and models. For those readers who want more technical detail or empirical evidence, each chapter includes a list of related readings. The chapters in this book provide a good place to begin researching or studying a topic in economics, and they also provide direction for where to go next.

To the extent possible, the chapters in the book follow a common format. They begin with a review of theory and then examine applications of the theory, relevant empirical evidence, policy implications, and future directions. This format reflects the typical approach of economists to a topic. They begin by asking what theory or models exist to help in understanding the behavior of the participants in decisions related to the topic. Participants may be consumers, producers, resource owners, agents of government bodies, or third parties who are affected by but not in control of the decisions made by other participants. Often, existing models will be sufficient for understanding the decision-making process, but if not, existing models will be modified or expanded or, occasionally, entirely new models may be developed.

The theoretical base is then applied to the decisions and behavior of participants relevant to the topic being explored. For example, an economist examining the decisions of owners of professional baseball teams may find that traditional models of profit maximization provide a good base but that they have to be modified to take into account motives that include status or pleasure in addition to profit. Whether existing or modified models are used, the economist's objective is to ask whether the theory or model can take into account the unique considerations critical to the topic.

Once the theory or model is developed, empirical evidence is explored, usually using statistical and econometric tools, to evaluate the ability of the model to predict outcomes. If the data lend support to the model, the model can then be used to predict outcomes. It is at this point that economic analysis leads to policy implications. Once economists have models that explain decision making and predict outcomes, policy makers have the basis for altering incentives to lead economic agents to make desirable choices. For example, once economists have identified the key variables influencing consumers' decisions about how much sugary soda to drink or whether or not to recycle soda cans, policy makers can establish or modify incentives for consumers to change their soda consumption and to recycle their cans instead of putting them in the trash.

The format of most chapters—theory, applications, empirical evidence, policy implications—is consistent with this common approach to economic analysis. Following the section on policy implications, most chapters discuss future directions—what are the new but related questions that are likely to be explored by economists; what new methods are being developed to analyze data on the topic; what insights from other disciplines are likely to be applied to this topic; what policies are likely to be developed related to the topic? Chapters in the book generally reflect this approach and the resulting format, but given the wide range of topics addressed, the format is not appropriate in every chapter. Some of the initial theory chapters, methodology chapters, and history chapters more logically follow a different structure, and common format has been sacrificed in favor of following the logic.

Identifying 100 economists to write chapters in their areas of expertise has been made easier by the assistance of the members of the editorial board. Jeff Ankrom from Wittenburg University, Robin Bartlett from Denison College, Karl Case from Wellesley College, and Wendy Stock from Montana State University provided contacts with economists from a wide range of universities and perspectives. In some cases, a potential author could not work this chapter into his or her writing schedule but provided information about young colleagues who were working in the same area. Some of the chapters dealing with more cutting-edge topics were written by economists identified in this roundabout way.

The members of the editorial board were also very helpful in identifying the list of topics that are represented in the 92 chapters of the handbook. And again, many good suggestions were made by authors who were contacted about writing on one topic but felt that the book would benefit from a new area of their research or from the work of a colleague. In a field as expansive as economics, these suggestions were invaluable.

Members of the editorial board also made valuable contributions to the book by reviewing many of the chapters and providing insights and suggestions for authors who, invariably, accepted recommendations with enthusiasm. The comments of other economists and policy makers who reviewed chapters have also contributed to ensuring that the book is correct, current, and balanced.

Jim Brace-Thompson, Acquisitions Editor, was an inspiring and reassuring voice from the inception of this project. Sanford Robinson, Developmental Editor, provided

needed advice, prompting, and encouragement throughout the writing process. I am very grateful to both of them.

Laura Notton and Leticia Gutierrez, SAGE Reference Systems Coordinators, are surely some of the most patient people I've worked with. They promptly answered innumerable questions and resolved the never-ending stream of technical problems associated with publishing a book with 100 authors.

Of course, the greatest thanks are owed to the authors of these chapters, many of whom struggled through multiple revisions to find just the right level of detail and depth of analysis. Explaining economics without calculus was a challenge for many of them, but they all succeeded in providing explanations that will be accessible to college students and other readers. Authors were also diligent about finding the most recent evidence and most interesting applications, writing right up to deadlines so that they could incorporate the latest government figures, conference proceedings, and journal articles. For senior economists whose contributions to this handbook reflect many years of work on a topic, I hope that it has been gratifying to write a chapter that is likely to provide crucial information and inspiration for many undergraduates as they embark on research projects. For younger contributors for whom this may be the first publication, I hope that this is just the first of many valuable contributions that will made to the discipline.

Working with 100 authors around the world meant that for the last few years I have been e-mailing, editing entries, and managing reviews during vacations, on holidays, and on weekends at all hours of the day and night. I am grateful to my children, Lindsey and Graham, who continued to be patient and understanding when I worked at times when I should not have and when I did not pay attention to the most important things in life.

*Rhona C. Free*
*Eastern Connecticut State University*

# About the Editor

**Rhona C. Free** (PhD, University of Notre Dame) is Vice President of Academic Affairs and a professor of economics at Eastern Connecticut State University. She was honored in 2004 as a Professor of the Year National Winner from The Carnegie Foundation for the Advancement of Teaching and the Council for Advancement and Support of Education. She was a founding member of the Connecticut Consortium for Learning and Teaching and of the Connecticut Campus Compact. In 2001, Free received ECSU's Distinguished Faculty Member Award.

Free is the author of many textbook supplements and has consulted frequently in the development of principles of economics textbooks. She has written on a range of topics, including teaching with technology and accommodating students with learning disabilities. Free's research interests focus on earnings differences by race and gender. Her current research analyzes effects of college major choice on differences by race and gender in expected starting salaries.

# About the Contributors

**Christopher Ross Bell** is an associate professor of economics at the University of North Carolina at Asheville. He received his BA from the University of California, Berkeley and his PhD in economics from the University of Pennsylvania. His main research areas are transaction cost economics, especially its intersection with political economy, and teaching economics as a laboratory science. He is also interested in Web programming and Web design and is taking advantage of those interests to create a Web application that will allow economics professors to display automatically updated economic data on their class Web sites.

**Nathan Berg** is an associate professor of economics in the School of Economic, Political, and Policy Sciences (EPPS) at University of Texas–Dallas. He has published more than 40 articles and chapters in behavioral economics since 2001 in journals such as *Journal of Economic Behavior and Organization, Social Choice and Welfare,* and *Contemporary Economic Policy.* Berg was a Fulbright Scholar in 2003, and his research has been cited in *BusinessWeek,* Canada's *National Post, The Village Voice, The Advocate,* and *Atlantic Monthly.* Berg spent 2005 at the Max Planck Institute–Berlin, Center for Adaptive Behavior and Cognition. In 2006, Berg was appointed to the editorial board of *Journal of Socio-Economics* and elected to the Board of the Society for the Advancement of Behavioral Economics. He sings and writes for the acoustic rock band Halliburton(s), and links to his work as a movie actor can be found at www.nathanberg.net.

**Mahadev Ganapati Bhat** is an associate professor of natural resource economics in the departments of earth and environment and economics at Florida International University. His current research focuses on the economics of coastal and marine resources, ecosystem restoration, water resources, and payment for environmental services. He is specialized in applying quantitative techniques—namely, spatiotemporal dynamic models, game-theoretic models, economic input-output models, and economic valuation techniques—to environmental problem solving. His research projects have been funded by the Everglades National Park, U.S. Department of Agriculture, U.S. Agency for International Development, the World Bank, the Government of India, and Ford Foundation. He has more than 150 publications, including refereed articles (30), conference publications, research reports, and book chapters. Dr. Bhat is a Berg Fellow of the Soil and Water Conservation Society. He routinely serves as a referee for several professional journals and is on the editorial board of the *Journal of Sustainable Agriculture*. Dr. Bhat received his PhD in agricultural economics (with specialization in natural resource economics) from the University of Tennessee in 1991 and a master's degree in the same area from the University of Agricultural Sciences, Dharwad, India, in 1983.

**Lyda S. Bigelow** is an assistant professor of strategy at the David Eccles School of Business, University of Utah. She received her PhD from the University of California, Berkeley. Her research focuses on using transaction cost economics to assess the impact of efficient boundary of the firm decisions (e.g., make-or-buy decisions, strategic alliances) on firm performance. Her most recent work has investigated the trade-offs of managing efficient sourcing arrangements under conditions of rapid technological innovation. Her work has appeared in the *Strategic Management Journal, Management Science,* and the *Journal of Economic and Organizational Behavior,* and she has won Best Paper awards from both the Entrepreneurship and Business Policy and Strategy Divisions of the Academy of Management. She is a member of the editorial board of the *Strategic Management Journal.* Prior to joining the faculty of the David Eccles School of Business at the University of Utah, she was a strategy professor at the John M. Olin School of Business, Washington University in St. Louis.

**Peter J. Boettke** is the BB&T Professor for the Study of Capitalism and University Professor of Economics at George Mason University (GMU). He received his PhD in economics from GMU in 1989. Prior to returning to GMU in 1998, Boettke held faculty positions at New York University and the Hoover Institution at Stanford University. Boettke's research is in the areas of comparative political and economic systems and their consequences with regard to material progress and political freedom, as well as twentieth-century economic thought and the methodology of the social sciences.

**Robert A. Boland** is a clinical assistant professor of sports management at New York University. He received his JD from Samford University's Cumberland School of Law and his AB degree from Columbia University. An admitted attorney and certified sports agent, his research interests

include collective bargaining, labor law, player compensation and antitrust law issues, and the history and social implication of sports.

**Michael E. Bradley** is a professor of economics at University of Maryland, Baltimore County. He received his AB degree from Albion College and his MS and PhD degrees from Cornell University. He has taught at Pennsylvania State University, Cornell University, the Paul H. Nitzke School of International Studies of the Johns Hopkins University, the U.S. Department of State Foreign Service Institute, and Colgate University. He has published articles and book chapters in the areas of comparative economics, international economics, and history of economic thought. He is the author of a two-volume text, *Microeconomics* (2nd ed.) and *Macroeconomics* (2nd ed.). His current teaching and research is in intermediate and graduate microeconomics, history of economics, and economic history of Russia and the USSR.

**Jennifer L. Brown** is an assistant professor in the department of economics at Eastern Connecticut State University. She earned her PhD in economics from the University of California, Santa Barbara. Her research focuses on environmental economics, energy economics, and industrial organization. Her most recent work evaluates the impact that environmental regulations have historically had on the petroleum industry. Additional work has focused on analyzing the impact that global climate change is expected to have on the economies of least developed countries.

**James D. Burnell** is a professor of economics and currently the chair of the urban studies program at the College of Wooster. He received his PhD from the University of Illinois in 1978. His research has recently focused on land use issues with particular consideration of the strategic interaction of the zoning decisions of suburban communities. His current research is on the neighborhood spillovers associated with housing tenure decisions and the availability of mortgage funds.

**Peter T. Calcagno** is an associate professor of economics in the department of economics and finance and director of the Initiative for Public Choice & Market Process at the College of Charleston. He received his PhD in economics from Auburn University. His research focuses on public choice and the political economy of voting institutions. He has published a number of articles in academic journals, including *Public Choice, Economics of Governance, Public Finance Review,* and the *Journal of Public Finance and Public Choice.*

**Neil Canaday** is currently a visiting assistant professor at Wesleyan University. His areas of interest include economic history, public finance, and labor economics. Related to public finance, he has publications examining tax assessment policy and discrimination in school resources during segregation.

**Ping Ching Winnie Chan** is a research economist in the analysis branch of Statistics Canada. She received her PhD from the University of Toronto in 2009. Her dissertation is on school choice and public school performance, and her research focuses on the economics of education, public economics, and labor economics.

**Carmel U. Chiswick** is a professor of economics (emeritus) at the University of Illinois at Chicago. She earned her PhD in economics from Columbia University in 1972, specializing in economic development, labor economics, and economic history. Her recent research focuses on the economics of religion, and she is a founding member of the Association for the Study of Religion, Economics and Culture (ASREC). A selection of her articles on Judaism from various journals and reports is now available in *The Economics of American Judaism* (2008).

**Victor V. Claar** is an associate professor of economics at Henderson State University, the public liberal arts college of Arkansas. He earned his PhD in economics at West Virginia University, where he wrote his doctoral dissertation under the guidance of Ronald Balvers. Prior to Henderson he held the rank of associate professor of economics (with tenure) at Hope College in Michigan, where he taught from 2000 to 2009. He spent a recent year as a Fulbright Scholar in Armenia giving graduate lectures and conducting research at the American University of Armenia. Claar is coauthor, with Robin J. Klay, of *Economics in Christian Perspective: Theory, Policy and Life Choices* (2007), and his scholarly articles have appeared in several peer-reviewed outlets, including *Applied Economics, Public Finance Review, and the Journal of Markets & Morality*. He recently completed work on a short book about fair trade, scheduled to be published by the Acton Institute in 2010.

**Paul F. Clark** is a professor and head of the department of labor studies and employment relations at Pennsylvania State University. He received his PhD in public policy and administration from the University of Pittsburgh. His research has focused on the structure and government of American unions and on collective bargaining in the coal, steel, and health care industries.

**Richard D. Coe** is an associate professor of economics at New College of Florida in Sarasota. He received a PhD in economics (1979) and a JD (1978) from the University of Michigan. His research has included issues in welfare use, the definition of poverty, and the legal requirements for a land tax.

**Richard R. Cornwall** is a professor emeritus of economics at Middlebury College in Middlebury, Vermont. He received a PhD in economics in 1968 from the University of California, Berkeley. His work has been in mathematical microeconomics and queer theory. Since 1998, he has been retired from teaching and living in San Francisco, devoting full time to research on the interaction between markets and social identities.

**Kenneth A. Couch** is an associate professor in the department of economics at the University of Connecticut and director of the Center for Population Research. He received his PhD and a master's degree from the University of Wisconsin and holds a master's degree from the University of Glasgow. His research interests include wage determination, disadvantaged groups in the labor market, and policy evaluation.

**Rosemary Thomas Cunningham** is the Hal and Julia T. Smith Professor of Free Enterprise in the department of economics at Agnes Scott College. Cunningham received her BA in mathematics/economics, MA in economics, and PhD in economics at Fordham University. Prior to coming to Agnes Scott College in 1985, she was an assistant professor at Fairfield University.

**Christopher S. Decker** is an associate professor of economics at the University of Nebraska at Omaha (UNO). He received his PhD in business economics from Indiana University's Kelley School of Business in 2000 and currently teaches managerial economics in UNO's MBA program. Decker has published numerous journal articles in the fields of environmental regulation, energy economics, and industrial organization. Recent work has focused on the market structure and workplace accident rates.

**George F. DeMartino** is an economist at the Josef Korbel School of International Studies, University of Denver. He has written extensively on economics and ethics, particularly in the context of international economic integration. He is the author of *Global Economy, Global Justice: Theoretical Objections and Policy Alternatives to Neoliberalism* (2000) and *The Economist's Oath: On the Need for and Content of Professional Economic Ethics* (in press).

**Gillian Doyle** (BA, Trinity College Dublin; PhD, University of Stirling) is the director of the Masters Programme in Media Management at the Centre for Cultural Policy Research at the University of Glasgow and is a visiting professor at the Institute of Media and Communications, University of Oslo. Her research interests are media economics and media and cultural policy, and she is currently President of the Association for Cultural Economics International (ACEI).

**Patricia A. Duffy** is Alumni Professor of Agricultural Economics and assistant provost for Undergraduate Studies at Auburn University. She received her PhD from Texas A&M University. She is the coauthor of *Farm Management* (6th ed.) and has authored or coauthored numerous journal articles in the areas of farm management and applied policy analysis. Her current research interests concern problems in farm management as well as the effect of nutrition programs on food security, diet quality, and obesity.

**Isaac Ehrlich** is State University of New York and University of Buffalo Distinguished Professor of Economics, Melvin H. Baker professor of American Enterprise, chair of the department of economics, and director of the Center of Excellence on Human Capital at the State University of New York at Buffalo. He is also a research associate at the National Bureau of Economic Research and editor-in-chief of the *Journal of Human Capital,* published by the University of Chicago Press. Ehrlich received his PhD from Columbia University and served as assistant and associate professor of business economics at the University of Chicago before joining SUNY Buffalo. He is one of the founders and leaders of the literature on the economics of crime and justice and has published extensively on other topics in leading journals of economics. He has been listed among the top 100 economists in published citations and has been included in all issues of *Who's Who in Economics: A Biographical Dictionary of Major Economists 1700–1980,* as well all of its later editions.

**Seda Ertaç** received her PhD in economics from the University of California, Los Angeles in 2006. She then joined the economics department of the University of Chicago as a postdoctoral scholar and is currently an assistant professor of economics at Koç University. Dr. Ertaç's fields of research are applied microeconomic theory and experimental economics. Her research agenda includes theoretically and experimentally studying the links between imperfect information and incentive systems in the context of effort and motivation in organizations, as well as the effects of gender and personality on economic behavior. Dr. Ertaç's experimental work to date has been supported by the U.S. National Science Foundation, Russell Sage Foundation, and the European Union.

**Erick Eschker** is a professor and chair of the department of economics at Humboldt State University, Arcata, California. He earned his PhD in economics from the University of California, Davis in 1997 and his bachelor's degree in economics from the University of Illinois, Urbana-Champaign. He was a research economist with the American Medical Association and is currently the director of the Humboldt Economic Index, which collects and analyzes data on the regional economy.

**Viktar Fedaseyeu** is currently a doctoral candidate in finance at Boston College. His primary research interests include bounded rationality, behavioral finance, corporate social responsibility, and the way financial markets react to uncertain information. He holds a degree in economics from Belarus State Economic University and was a student at Eastern Connecticut State University.

**Ann Harper Fender** is a professor of economics (emeritus) at Gettysburg College, Gettysburg, Pennsylvania. She received her PhD from Johns Hopkins University. Her research involves studies of the economics of the fur trade and contemporary issues relating to the structure of the telecommunications industry.

**Aju Fenn** is the John L. Knight Professor of Free Enterprise and chair of the department of economics and

business at Colorado College in Colorado Springs, Colorado. Past honors include the Lloyd E. Worner Teacher of the Year award and the John D. and Catherine T. MacArthur professorship. He teaches the economics of sports, mathematical economics, intermediate microeconomic theory, research methods and econometrics. He also specializes in the economics of addiction. His work in sports economics focuses on the measurement of competitive balance in sports and the determinants of competitive balance. He has also published work on the value of a sports team to an area. He has served as a referee for journals such as *Economics Letters, Southern Economic Journal, Contemporary Economic Policy, International Journal of Sport Finance,* and *Journal of Sports Economics.*

**Satyananda J. Gabriel** is a professor and chair in the department of economics at Mount Holyoke College in South Hadley, Massachusetts, and academic coordinator of the Rural Development Leadership Conference Summer Institute at the University of California, Davis. He received his PhD from the University of Massachusetts at Amherst. Gabriel is the author of *Chinese Capitalism and the Modernist Vision* (2006) and *Rising Technomass: The Political Economy of Social Transformation in Cyberspace* (2010). His current research interests include Chinese and East Asian economic development, corporate finance, and comparative economic systems.

**Thomas Gall** is an assistant professor of economics at the University of Bonn in Germany. He obtained his diploma in 2001 after his undergraduate studies in economics at the University of Munich, Germany, and Pompeu Fabra University in Barcelona, Spain. In 2005, he completed a PhD in economics after graduate studies at Mannheim University, Germany, and University College London, UK. His research interests lie in microeconomic theory, and he has published on imperfect matching markets, occupational choice, and development economics.

**Lawrence D. Gwinn** is an associate professor of Economics and Chair of the department of economics at Wittenberg University, Springfield, Ohio. He received his PhD from the University of Kansas. His research interests include the international transmission of economic disturbances and monetary policy effectiveness under alternate exchange rate systems.

**Thomas W. Harvey,** DBA, is an associate professor of finance and the director of the Institute for Contemporary Financial Studies at Ashland (Ohio) University, where he is responsible for the asset management track of the finance curriculum and for directing various student research initiatives. Dr. Harvey received his doctorate in management strategy and international business from Cleveland State University. His MBA is in finance from Case Western Reserve University, with his BA in English from Hillsdale College. Dr. Harvey's research interest is behavioral finance and the changing nature of investor behavior. He is also a student of the financial markets in the United States and has published two books that focus on the commercial banking industry: *Quality Value Banking*, with Janet L. Gray, and *The Banking Revolution*. Prior to joining the faculty at Ashland, Dr. Harvey's career was spent in the commercial banking industry.

**Sue Headlee** is an associate professor of economics at American University, Washington, D.C., where she teaches the Washington Semester Program in Economic Policy. She received her PhD in economics at American University. She is the author of three books: *The Political Economy of the Family Farm: The Agrarian Roots of American Capitalism; The Cost of Being Female,* coauthored with Margery Elfin; and *A Year Inside the Beltway: Making Economic Policy in Washington.* She is the author of two articles: "Income and Wealth Transfer Effects of Discrimination in Employment: The Case of African Americans, 1972–1990," in *The Review of Black Political Economy,* and "Economic History, Western Europe," in the *Elgar Companion to Feminist Economics.*

**Gillian Hewitson** is in the political economy department at the University of Sydney. She received her PhD in economics and women's studies from La Trobe University in Melbourne, Australia. She is the author of *Feminist Economics* (1999) and journal articles and book chapters in the areas of feminist economics and monetary theory. Her current research interests include gender, race, and class in the history of economic thought, particularly in relation to Australia, and the political economy of the food supply.

**Michael Hillard** is a professor of economics at the University of Southern Maine. He has published widely in the fields of labor relations, labor history, and the political economy of labor in academic journals, including *Labor: Studies in the Working Class Histories of the Americas, Labor History, Review of Radical Political Economics, Advances in Industrial and Labor Relations, Journal of Economic Issues, Historical Studies in Industrial Relations,* and *Rethinking Marxism.* He has coauthored many articles on a Marxian analysis of industrial relations with Richard McIntyre, professor of economics at the University of Rhode Island. His essay titled "Labor at Mother Warren" won Labor History's "Best Essay, U.S. Topic" prize for 2004.

**Steven Horwitz** is the Charles A. Dana Professor of Economics at St. Lawrence University in Canton, New York. His PhD is from George Mason University, and he has written extensively on the Austrian School of Economics, macroeconomics, and political economy, as well as recent work on the economics and social theory of the family. He is currently completing a book manuscript on classical liberalism and the evolution of the western family.

**David Hudgins,** lecturing professor of economics at the University of Oklahoma, received his PhD in 1993 from the University of Illinois at Urbana-Champaign. He is also

a Certified Financial Manager and has served as a financial adviser for Merrill Lynch. His recent publications include articles in *Computational Economics* and *Review of Applied Economics.*

**Shawn Humphrey** is an assistant professor of economics at the University of Mary Washington in Fredericksburg, Virginia. He earned his PhD from Washington University in Saint Louis. His research is concerned with uncovering the political-military foundations of economic prosperity.

**Steven L. Husted** is a professor of economics at the University of Pittsburgh. Professor Husted has published widely in the areas of international trade, international finance, and monetary economics and is coauthor with Michael Melvin of a popular textbook on international economics. In 1986–1987, he spent a year as a senior staff economist specializing on trade policy issues on the President's Council of Economic Advisers. He has had wide-ranging international experience, including visiting appointments to the Australian National University, the University of Glasgow, and the University of Strathclyde. His current research interests include studies of the growth and geographic extent of Chinese exports, financial capital flows to developed economies, and long-run exchange rate behavior.

**Lee H. Igel** is a clinical assistant professor of sports management at New York University. He received his PhD in industrial/organizational psychology from Capella University and holds degrees in both counseling and clinical exercise physiology from Boston University. A frequent writer and speaker on human affairs, his areas of study are in management and organizations.

**Elizabeth J. Jensen** is a professor of economics at Hamilton College in Clinton, New York, where she has taught since 1983. Professor Jensen received a BA in economics from Swarthmore College and a PhD in economics from the Massachusetts Institute of Technology. Before coming to Hamilton, she worked at the Council of Economic Advisers. Professor Jensen has been honored for her teaching, receiving a prestigious award for outstanding teaching from Hamilton College. Jensen is coauthor of *Industrial Organization: Theory and Practice,* a leading industrial organization textbook developed in part from experiences teaching students at Hamilton College. Her recent work investigates the predictors of academic success in college, student course choice, and the determinants of students' interest in economics. Jensen teaches courses in industrial organization, antitrust and regulation, American economic history, and microeconomic theory.

**Shirley Johnson-Lans** is a professor and the chair of the department of economics, Vassar College. A labor economist who teaches and does research in the areas of health economics, economics of education, and gender studies, she is the author of *A Health Economics Primer* (2006) and of numerous articles and working papers. She is a member of the City University of New York/Columbia University faculty seminar on demography and health economics and of the New York Health Policy Group. She holds a PhD in economics from Columbia University; an MA in political economy from Edinburgh University, where she was a Marshall Scholar; and a BA magna cum laude in philosophy from Harvard.

**Nicholas A. Jolly** is a visiting assistant professor in the department of economics at Central Michigan University. He received his master's and PhD in economics from the University of Connecticut. Prior to joining the faculty at Central Michigan University, he worked as an economist at the Connecticut Department of Labor. His research interests include wage determination, job displacement, applied microeconomics, and policy analysis.

**Rebecca P. Judge** is an associate professor of economics and environmental studies at St. Olaf College in Northfield, Minnesota. After receiving her MS in biology from the University of Minnesota, Duluth, and her PhD in economics from Duke University, she has spent her career engaged in questions relating to the intersection of economics and the environment. Her publications include a study exploring least-cost methods for implementing an endangered species preservation constraint on public lands, an analysis of household response to alternative incentives to engage in recycling, and an exploration of the property regimes influencing John Locke in the development of his theory of property.

**Fadhel Kaboub** is an assistant professor of economics at Denison University (Granville, Ohio) and research associate at the Levy Economics Institute of Bard College (New York), the Center for Full Employment and Price Stability (Missouri), and the International Economic Policy Institute (Ontario, Canada). He taught at Drew University, where he was also co-director of the Wall Street Semester Program; the University of Missouri–Kansas City (UMKC); and Bard College at Simon's Rock. Kaboub's research is in the post-Keynesian and institutionalist tradition in the fields of macroeconomic theory and policy, monetary theory and policy, and economic development. His work has been published in the *Journal of Economic Issues, Review of Radical Political Economics, Review of Social Economy, International Journal of Political Economy,* and *International Labour Review*. He has been a member of the editorial board of the *Review of Radical Political Economics* since 2006 and has been the book review editor of the *Heterodox Economics Newsletter* since 2007. He holds a PhD in economics from UMKC.

**Bradley P. Kamp** is an associate professor of economics at the University of South Florida. His areas of interest include product quality, predatory pricing, and signaling models. His work has appeared in journals such as *International Journal of Industrial Organization, Southern Economic Journal, Economic Inquiry,* and *Review of Law and Economics.* He earned a BA from the University of Illinois and a PhD from the University of California, San Diego.

**Pavel S. Kapinos** is an assistant professor of economics at Carleton College in Minnesota. He received his PhD in economics from the University of Illinois at Urbana-Champaign. His research focuses on the optimal design of monetary policy.

**Stephen H. Karlson** is an associate professor of economics at Northern Illinois University. He completed a PhD in economics from the University of Wisconsin at Madison in 1980. His research interests have included regulation of electric utilities, technology adoption in steelmaking, and irreversible investments. He is currently an Environmental and Energy Policy Scientist with Argonne National Laboratories on a temporary basis.

**Peter Kennedy** is professor emeritus at Simon Fraser University, Canada. He received his BA from Queen's and his PhD from Wisconsin. He has published widely in economic education and in econometrics and is best known for his book *A Guide to Econometrics,* perhaps the best single source on the full range of econometric topics.

**Markus Kitzmueller** is an Erb Postdoctoral Research Fellow at the S. M. Ross School of Business/SNRE of the University of Michigan (Ann Arbor). Currently, he is expecting his PhD in economics as a Researcher of the European University Institute (EUI) in Florence (Italy). He has graduated magna cum laude with an MA in European Economic Studies from the College of Europe in Bruges (Belgium) and has worked for the European Commission (DG ECFIN) in 2004–2005. His research focus is on applied microeconomics, especially information economics, contract and organization theory, industrial organization, as well as public economics. Current interests center on the interaction between strategic firm behavior and public policy in light of corporate social responsibility.

**Kevin C. Klein** is a professor of economics at Illinois College in Jacksonville, Illinois. Dr. Klein earned a Doctorate of Arts in Economic Education at Illinois State University. His teaching focus is economics at the principles and intermediate undergraduate levels with special interest in environmental economics. He has authored or coauthored several teaching manuals, test banks, and study guides. He is also a coauthor of a survey of economics textbook.

**Agneta Kruse** is a senior lecturer in economics at Lund University in Sweden. Her research focuses on economics of social insurance and pension systems. She served as an expert to the parliamentary Commission on Pensions preceding the Swedish pension reform. She analyzes, among other things, pay-as-you-go pension systems and their sustainability encountering demographic and economic changes as well as the political economy of pension reforms. She has published articles in journals and contributed chapters to a number of books. Lately, she has also been working on globalization and social insurance.

**Roy Love** obtained his PhD at the University of Leeds (UK). He has lectured in economics at the universities of Botswana, Lesotho, and Addis Ababa and at Sheffield Hallam University, England. He has publications on the subject of HIV/AIDS and is currently an independent researcher and consultant with experience of economic evaluation for major international donors on HIV/AIDS and health projects in Africa.

**Yahya M. Madra** teaches political economy and history of economics at Gettysburg College. He is also an associate editor of *Rethinking Marxism: A Journal of Society, Economics and Culture.* He has published on the methodology and philosophy of economics, as well as the intersection between Marxian political economy and Lacanian psychoanalysis. His writings have appeared in the *Journal of Economic Issues, Rethinking Marxism, Psychoanalysis, Society and Culture,* and edited volumes. Currently, he is working on the intellectual genealogy of neoliberalism and its variants.

**Leah Greden Mathews** earned her PhD in agricultural and applied economics from the University of Minnesota and is associate professor of Economics at the University of North Carolina at Asheville. Her research as an environmental economist has focused on nonmarket valuation and the links between economics and policy. Dr. Mathews's current research, The Farmland Values Project, estimates the values that communities in western North Carolina have for farmland, with particular attention to those values that are not typically exchanged in markets such as scenic beauty and cultural heritage.

**Sandra Maximiano** is an assistant professor of economics at Krannert School of Management, Purdue University. After receiving her PhD from the University of Amsterdam, in 2007 she started as a postdoctoral scholar in the economics department of the University of Chicago. Dr. Maximiano's research interests lie in the fields of behavioral and experimental economics, labor economics, economics of education, and organizational economics. Her projects span various issues and are built on both laboratory and field experiments and microeconometric tools to investigate questions related to social preferences and reciprocity, education and training, incentive systems, and gender and culture differences in economic decisions.

**Ken McCormick** is a professor of economics at the University of Northern Iowa. He was an early critic of the AD/AS model. McCormick has published numerous papers on a variety of topics, including the AD/AS model. His book *Veblen in Plain English* (2006) has been well received and led to an appearance on the *Bob Edwards Show.*

**Rachel McCulloch** is the Rosen Family Professor of International Finance at Brandeis University. Prior to her appointment at Brandeis, she was a faculty member at the University of Chicago, Harvard University, and the University of Wisconsin–Madison. She received her PhD in economics from the University of Chicago in 1973. Her current research focuses on international economic policy.

**John D. Messier** is an assistant professor of economics at the University of Maine at Farmington. He earned his PhD from American University and studies international development issues. Dr. Messier spent 5 months in Ecuador working with informal workers on credit issues and the impact of financial shocks on well-being. Currently, Dr. Messier is working on the impact of fair trade on income and nutrition for coffee producers in Nicaragua.

**Laurence Miners** is the director of the Center for Academic Excellence and a professor of economics at Fairfield University in Fairfield, Connecticut. He earned his doctorate in economics at the University of North Carolina at Chapel Hill. His research interests focus primarily on economic pedagogy and course design. He has led faculty development workshops at other colleges and universities and made presentations at regional and national conferences focusing on both economics and pedagogy.

**Michael R. Montgomery** is an associate professor of economics at the University of Maine. He received his PhD from the University of Florida. He works in macroeconomics, monetary theory, and public economics. He has done work on the history of macroeconomics as well as applied work on the macroeconomic implications of time-intensive capital production periods and complementarity between types of capital goods.

**Mehdi Mostaghimi** is a faculty member in the school of business at Southern Connecticut State University. He is the author of many journal articles, book chapters, and technical reports on economic turning point forecasting, combining forecasts, consensus and group decision making, and strategic decision making.

**Shannon Mudd** has ventured between academics and nonacademics in his career as an economist. While completing a PhD in economics at the University of Chicago, he spent 2 years working as an analyst for the Federal Reserve Bank of Atlanta and taught for another year at Comenius University in Bratislava, Slovakia. After finishing his degree, he spent 4 years working on a USAID project in Moscow on taxation and intergovernmental relations. He has taught at the MA, MBA, and undergraduate levels and is currently an assistant professor at Ursinus College. His research has most recently focused on issues of access to finance for small- and medium-sized enterprises and the effect of banking crises on future expectations of banking crises.

**Sean E. Mulholland** is an associate professor of economics at Stonehill College. He received his BS and MA in economics from Clemson University in 1997 and 2001, respectively. He earned his PhD in applied economics from Clemson University in 2004. His research interests include the long-run economic growth within the United States, the economics of race and religion, and private and public land conservation.

**Kathryn Nantz** is an associate professor of economics at Fairfield University in Fairfield, Connecticut. She earned her PhD from Purdue University. Her teaching and research work is primarily at the intersections of the fields of labor economics and comparative economic systems. She is interested in how various incentive structures affect the behavior of workers in their jobs and in how political and cultural norms can create widely varying labor market outcomes in different settings.

**Lena Nekby** received her PhD from Stockholm University and is now an assistant professor at the department of economics, Stockholm University, and affiliated with the Stockholm University Linnaeus Center for Integration Studies (SULCIS). She is also an IZA Research Fellow. Her research is focused primarily on labor market issues relating to ethnicity, migration, and gender.

**Christopher J. Niggle** received his BA from Arizona State University (1967), MA from New School University (1970), and PhD in economics from the University of California, Riverside (1984). Since 1983, he has taught at the University of Redlands in Southern California, where his teaching responsibilities include courses in macroeconomics, money and banking, history of economic thought, and comparative economic systems. His research and publications have primarily been in the areas of money and macroeconomics. Current interests include comparisons of institutionalist and post-Keynesian macroeconomics with New Keynesian macroeconomics.

**Peter Nyberg** is an assistant professor at the Helsinki School of Economics. He obtained his PhD from the Hanken School of Economics in Helsinki. His research interests include asset pricing and financial econometrics.

**Lindsay Oldenski** is assistant professor of economics at the School of Foreign Service at Georgetown University. She received her PhD in economics from the University of California at San Diego, master's degree in public policy from the Kennedy School of Government at Harvard University, and BA from Guilford College. She has taught international trade and microeconomics at the Johns Hopkins University School of Advanced International Studies and at California State University, San Marcos. Her research interests include international trade in services and the organization of multinational activities.

**Anita Alves Pena** is an assistant professor of economics at Colorado State University. She received her PhD in economics from Stanford University in 2007, her MA in economics from Stanford University in 2004, and her BA in economics from the Johns Hopkins University in 2001. Her research interests are in public sector economics, economic development, and labor economics, and her current research relates to undocumented and documented immigration, public policy, poverty, and agricultural labor markets.

**Dante Monique Pirouz** is an assistant professor at the Ivey School of Business at the University of Western Ontario. She earned her PhD at the Paul Merage School of Business at the University of California, Irvine. Her

research interests include neuroeconomics and consumer decision making. Her current work focuses on the neural response of addictive product users such as cigarette smokers to marketing cues. She is a trained researcher in functional magnetic resonance imaging (fMRI) and has received the Athinoula A. Martinos Center for Biomedical Imaging Functional MRI Visiting Fellowship at Harvard Medical School and Massachusetts General Hospital. She also has an MBA from the Wharton School of Business and an MA from the Lauder Institute at the University of Pennsylvania, and she graduated cum laude with a BA from the University of California, Los Angeles.

**Clifford S. Poirot Jr.** is currently an associate professor of economics in the department of social sciences at Shawnee State University, where he teaches both standard economics courses as well as courses that focus on sociocultural evolution. He has also been the director of the Shawnee State University Honors Program and is currently president of the Faculty Senate. He completed his BS in economics and political science in 1984 at Guilford College, Greensboro, North Carolina, where he wrote an undergraduate thesis on the evolution of political violence in Guatemala. He received his PhD in economics in 1991 from the University of Utah, where he wrote his dissertation on the evolution of the open field system in late medieval and early modern England. He has also taught at Eastern Washington University in Spokane, Washington; the University of Timisoara in Timisoara, Romania, under the auspices of the Civic Education Project; the American University in Bulgaria, in Blagoevgrad; and Mary Washington College in Fredericksburg, Virginia. He is the author of multiple articles on sociocultural evolution and related topics and has published in the *Journal of Economic Issues, Journal of Post-Keynesian Economics,* and *The Forum for Social Economics.*

**Indrajit Ray** is a chaired professor in the department of economics, University of Birmingham, and currently leads the economic theory group in his department. His area of academic research is game theory, general equilibrium theory, experimental economics, and environmental economics. He was trained in premier institutions such as the Indian Statistical Institute, India, and CORE, Belgium. He has taught at University of York and Brown University. Professor Ray has published numerous research articles in international journals, including in premier journals in his field such as: *Games and Economic Behavior*, *Economic Theory*, *Journal of Mathematical Economics*, *Social Choice and Welfare*. He has served as an editor of the journal *Bulletin of Economic Research* during 2000–2005 and currently is the associate editor of the electronic open-access journal *Quantitative and Qualitative Analysis in Social Sciences*. He has visited different universities in several countries to present his research in workshops and seminars. Professor Ray is also an active educationist and the chairman of a trust called *Vidyapith* that supports education in India.

**Marta Reynal-Querol** is an ICREA Research Professor at the department of economics and business at Pompeu Fabra University. She is also an affiliated professor at the Barcelona Graduate School of Economics. She is a member of the editorial board of the *Journal of Conflict Resolution* and the *European Journal of Political Economy* and is associate editor of the *Spanish Economic Review*. She is an award holder of the ERC-starting grant, awarded by the European Research Council. She holds a PhD in economics from the London School of Economics and Political Science (2001) and Master with Honors from Pompeu Fabra University. Reynal-Querol worked at the World Bank between 2001 and 2005. Her research has been concentrated on the causes of civil wars and genocides, conflict resolution and the aftermath of conflict, aid effectiveness, and the economics of institutions. She has published in *American Economic Review, Review of Economics and Statistics, Economic Journal, Journal of Economic Growth, Journal of Development Economics, European Journal of Political Economy, Journal of Conflict Resolution, Journal of Comparative Economics, Defence and Peace Economics,* and *Economic Letters,* among others.

**Anthony M. Rufolo** is a professor of urban studies and planning at Portland State University, where he specializes in state and local finance, transportation, urban economics, and regional economic development. He has a BS in economics from the Massachusetts Institute of Technolocy and a PhD in economics from the University of California, Los Angeles. Prior to joining the faculty at Portland State in 1980, he spent 6 years as Economist and Senior Economist with the Federal Reserve Bank of Philadelphia. Dr. Rufolo has extensive experience in transportation research and policy, including analyses of articulated buses, weight-mile taxes for heavy vehicles, the effect of road capacity on the time distribution of travel, the effect of access to light rail on home values, and the effects of pricing on miles driven. Dr. Rufolo has served on the Economics Committee of the Transportation Research Board and has served as a Visiting Scholar at the U.S. Department of Transportation.

**Benjamin Russo** teaches economics at the University of North Carolina at Charlotte. He received a PhD in economics from the University of Iowa. His recent research studies the effects of taxation on economic growth and state and local taxes. Russo has served on a number of state tax study commissions and as a tax consultant to Department of Finance Canada.

**Julie Sadler** is an assistant professor, department of labor studies and employment relations, Penn State University. She earned her doctorate and master's degree from the School of Industrial and Labor Relations at Cornell University. Her research focuses on the internal dynamics of labor unions in the United States, with a particular emphasis on leadership development and member engagement labor unions and nonprofits.

**Steven J. Shulman** is a professor of economics at Colorado State University. He received his PhD, MA, and BA from the University of Massachusetts at Amherst. He is the editor of *The Impact of Immigration on African Americans* (2004) and the author of numerous scholarly articles on the economics of racial inequality, low-wage work, and immigration.

**Brian Snowdon** is currently a senior teaching fellow (part-time) at the department of economics and finance at Durham University, UK. Previously, before retiring, he was professor of economics at Newcastle Business School, Northumbria University. He received his undergraduate and postgraduate degrees in economics from Leicester University. His main research interests are in the areas of macroeconomics and international growth and development. As well as numerous articles in academic journals, he has authored/coauthored/coedited 10 books in the area of economics, including *Conversations on Growth, Stability and Trade* (2002), and *Modern Macroeconomics: Its Origins, Development and Current State* (with H. R. Vane, 2002). His most recent book, *Globalisation, Growth and Development: Conversations With Eminent Economists,* was published in 2007.

**Maritza Sotomayor** is a lecturer at Weber State University, Utah. She received her master's degree in economics in 1990 from Centro de Investigación y Docencias Económicas (CIDE), Mexico City, and her PhD in applied economics in 2008 from Universidad Autónoma de Barcelona, Spain. She has published chapters in books and coauthored journal articles. Her research interest includes intra-industry trade, *maquiladoras,* and Latin America's external trade.

**Lawrence M. Spizman** is professor of economics at the State University of New York at Oswego. He received his PhD in economics in 1977 from the State University of New York at Albany. He has been a practicing forensic economist since 1985. He has published 27 articles, with many of them in the field of forensic economics. Dr. Spizman has coauthored the seminal work on estimating the damages to a minor child, "Loss of Future Income in the Case of a Personal Injury to a Child: Parental Influence on a Child's Future Earnings." He is called to consult throughout the United States.

**Mikael Svensson** is an associate senior lecturer in economics at the Swedish Business School, Örebro University, Sweden. He is also an affiliated researcher at the Centre for Research on Child and Adolescent Mental Health at Karlstad University, Sweden. He received his PhD from the Swedish Business School at Örebro University in 2007. His current research interests include health economics and economic evaluation.

**Leila Simona Talani** joined the department of European studies of King's College London in 2009 as lecturer in International and European Political Economy. She was previously a lecturer in European Politics at the University of Bath and a research fellow and then lecturer at the European Research Institute of the London School of Economics. In 2001 she spent a year as associate expert on migration issues at the United Nations Office for Drug Control and Crime Prevention in Cairo. She gained a PhD with Distinction from the European University Institute in Florence in 1998. Her thesis has been published as Betting for and Against *EMU: Who Wins and Who Loses in Italy and the UK From the Process of European Monetary Integration* (2000). She is also author of *European Political Economy: Political Science Perspectives* (2004), *Between Growth and Stability: The Demise and Reform of the Stability and Growth Pact* (2008), *Back to Maastricht* (2008), *EU and the Balkans: Policy of Integration and Disintegration* (2008), *The Future of EMU* (2010), *From Egypt to Europe* (2010), and *The Global Crash* (in press). Her current research interests focus on the political economy of migration flows from southern Mediterranean countries to the EU and on the credibility of exchange rate commitments and economic agreements.

**Troy L. Tassier** is an associate professor of economics at Fordham University in New York City. Prior to his position at Fordham, he was a postdoctoral research fellow at the Center for the Study of Complex Systems at the University of Michigan. He received a PhD in economics from the University of Iowa in 2002. Dr. Tassier's research focuses on complex systems research in economics and particularly diffusion processes in social networks. He has published papers on the effects of referral hiring and social network structure on labor market inequality and segregation, the evolution of social networks to optimize job information flows, and how the structure of social networks influences the spread of fads and fashions. His current research continues the study of job information networks as well as additional topics such as the spread of infectious diseases across social networks, how the structure of firms and other organizations influences problem-solving capabilities, and mechanisms of public goods provision.

**Peter Skogman Thoursie** received his PhD from Stockholm University and is now an associate professor at the Institute for Labour Market Policy Evaluation (IFAU), Uppsala in Sweden. His research deals primarily with empirical labor economics.

**Frederick G. Tiffany** is an associate professor of economics at Wittenberg University in Springfield, Ohio. He received his BA in economics from Kenyon College (1977) and PhD in economics from the University of Pennsylvania (1988). He teaches intermediate microeconomic theory, game theory, industrial organization, managerial economics, mathematics for economists, and public finance. His current research interest is the market for college education, especially the use of price discrimination by monopolistically competitive colleges.

**Kiril Tochkov** holds an MA in Chinese studies from the University of Heidelberg, Germany, and an MA and PhD in economics from the State University of New York at

Binghamton. He has studied at the Beijing Foreign Studies University in China and has worked as assistant manager in the marketing division of Volkswagen Automotive Co. Ltd. in Shanghai. He is currently an assistant professor of economics at Texas Christian University in Fort Worth, Texas, where he regularly teaches a class in Asian economics. His research focuses on regional growth, convergence, and efficiency in China.

**Paola Tubaro** earned a PhD in economics at the University of Paris-Ouest and the University of Frankfurt. She is currently a lecturer at the University of Greenwich, Business School in London and an associate member of Centre Maurice Halbwachs in Paris. Among her research interests are the history of economic thought and the philosophy and methodology of economics, with focus on the history of mathematical modeling and the methodology of experimental and computational economics.

**John Vahaly** is chair of the department of economics at the University of Louisville. He attended Vanderbilt University as a graduate student. He has authored teaching materials for principles texts in macroeconomics and has published articles about the macroeconomics of emerging market economies. His approach to teaching macroeconomics is primarily historical, showing how macroeconomics has changed over time. This presents students with a better understanding of current problems and what new challenges they present.

**Carlos Vargas-Silva** is an assistant professor of economics at Sam Houston State University. His research interests are workers' remittances, exchange rates, monetary policy, and time-series econometrics. He holds a PhD in applied economics from Western Michigan University.

**Douglas M. Walker** is an associate professor of economics at the College of Charleston, in Charleston, South Carolina. He received his PhD in economics from Auburn University. His research focuses on the economic and social effects of legalized gambling, especially casinos. Walker has published a number of articles in academic journals, and his book, *The Economics of Casino Gambling,* was published in 2007.

**Douglas O. Walker** is professor of economics in the Robertson School of Government at Regent University. He holds a PhD from the University of Southern California. Before joining Regent, Dr. Walker was a senior economist with the United Nations Secretariat, where he drafted studies of trends and prospects for the world economy and served as chief of quantitative research, secretary of the United Nations systemwide committee on technical research, and speechwriter to the Under-Secretary-General. While with the secretariat, Dr. Walker was adjunct professor of economics at Baruch College of the City University of New York and a member of the New York Academy of Sciences.

**Quan Wen** is a professor of economics at Vanderbilt University. He received his PhD in economics from the University of Western Ontario in Canada. Wen has authored and coauthored numerous journal articles on game theory and applied game theory in economics. His current research interests include repeated game, bargaining theory, mechanism design, and industrial origination.

**Sarah E. West** is an associate professor of economics at Macalester College in St. Paul, Minnesota. She received a PhD in economics from the University of Texas at Austin. Her research analyzes optimal tax policy, focusing on behavioral responses to policies for the control of vehicle pollution, including taxes on gasoline, subsidies to clean vehicles, and Corporate Average Fuel Economy standards. From 2002 to 2007, she was a member of the National Bureau of Economic Research's Working Group in Environmental Economics. Her work has been published in *American Economic Review, Journal of Environmental Economics and Management, Journal of Public Economics, Journal of Transport Economics and Policy, National Tax Journal,* and *Regional Science and Urban Economics.*

**David Wiczer** is a doctoral candidate at the University of Minnesota and research assistant at the Federal Reserve Bank of Minneapolis. He received his master's from the University of Illinois, Urbana-Champaign. He studies macroeconomics, with work ranging from optimal control in growth models to currency depreciation in sovereign default. His current focus is on computational methods and using micro-level data to estimate macroeconomic models.

**Amy M. Wolaver** is an associate professor of economics at Bucknell University. She received her PhD from the University of Wisconsin–Madison in 1998. Her research areas focus on the impact of public provision of family planning coverage on contraceptive use, pregnancies, pregnancy outcomes, substance use, and labor market effects of internal migration. She teaches courses in introductory economics, microeconomic theory, and health economics, as well as a team-taught course on HIV/AIDS.

**Timothy A. Wunder** received his PhD from Colorado State University in the fall of 2003. His dissertation was a history of thought on the origins of economic sociology, emphasizing the contributions from both Thorstein Veblen and Joseph Schumpeter. Since receiving his PhD, he has written several pieces on the history of economic thought and on economic education. Dr. Wunder currently is Senior Lecturer in the department of Economics at the University of Texas at Arlington.

# PART I

## Scope and Methodology of Economics

# 1

# History of Economic Thought

Paola Tubaro

*University of Greenwich; Centre Maurice Halbwachs*

Like economic history, the history of economic thought (HET) investigates economic issues in long-run perspective. Yet the two fields are distinct and should not be confounded: HET does not study economic *facts* such as the 1929 financial crisis but rather economic *theories* and economic *literature.* It focuses on the historical roots of economic ideas and takes into account a wide array of schools of thought; in conjunction with pure theory and the history of facts, it constitutes part of a comprehensive approach to the study of the economic bases of modern societies. Examples of research questions in HET include how economic issues (say, unemployment or growth) have been dealt with at different points in time, what intellectual debates they have stimulated, and what solutions have been concocted; how the meaning and interpretation of economic concepts such as involuntary unemployment or market equilibrium may have changed over time and how they have affected policy debates; and whether and how exchanges with neighboring disciplines or with currents of thought in philosophy have exerted an impact on the development of economics.

A variety of approaches to the study of HET coexist, but it is possible to identify two broad tendencies into which most of them would fit: one more "theory oriented," the other more "history oriented." The theory-oriented approach, often referred to as *history of analysis,* emphasizes *continuity* between present and past reflection, so that earlier writers are seen as a source of inspiration that may assist today's researchers in devising new solutions to current theoretical questions. Because economics is an approach to the study of society that endeavors to look beneath context-specific factors to discover underlying regularities in the behavior of individuals and communities, older economists who have already identified some of these regularities can provide useful insight despite the time distance that separates them from us. Earlier ideas have remained partly underdeveloped, and getting back to them can potentially suggest new directions for research. For some scholars, this means reviving alternative explanations and interpretations of economic phenomena, in a critical perspective with respect to any consensus that might exist today. In line with this methodological stance, historians of analysis primarily rely on the conceptual tools of contemporary economic theory.

In contrast, the history-oriented approach emphasizes *discontinuity* between different stages of development in economics and the specificities of each of them. The idea is that an economic theory is embedded in its historical, sociopolitical, and institutional environment and constitutes a response to the problems of the day, so that it is incorrect to understand it in abstraction from them. Specifically, a "retrospective" interpretation of past economic theories in light of the knowledge that has been subsequently acquired is impoverishing because it tends to present them all as imperfect, preliminary drafts of today's supposedly superior models. Instead, it is essential to detail the context in which an older theory emerged, and useful insight may come from historical techniques such as archival work, biographies, and oral history. This approach emphasizes the relative character of economic theories and their connections to politics and society at large. Among scholars who work along these lines, a large group has developed a close association with, and adopted methods of, science studies, in some cases with a critical perspective toward economics.

While the two above-outlined approaches are the object of recurrent and sometimes lively controversies within the HET community, many scholars are in fact aware that each has both strengths and weaknesses and try to combine the two in their research. Still other approaches may be occasionally present, particularly in the case of research at the crossroads between HET and the philosophy or methodology of economics.[1]

What follows is a brief overview of the main areas of research in HET that also correspond to a classic division of phases of development of the economics discipline from its origins to its present state. Following the *Journal of Economic Literature* classification, the evolution of economics has been subdivided into two phases—namely, HET through 1925 and HET since 1925; a shorter third part on recent developments has been added to this basic scheme.

There is obviously insufficient room to cover *all* aspects of the intellectual reflection in economics over such a long time span. For this reason, the presentation is limited to a sketch of what each period contributed to the study of three foundational issues in economics—namely, the theory of individual economic behavior, the market mechanism as a coordinating device, and the respective roles of markets and governments in the regulation of economic systems. Reflection on these issues has progressively formed economists' understanding of society and presently allows applications to a broad range of social phenomena, from monetary and financial matters to health and the environment. Furthermore, these very issues have been the object of major controversies that have divided economists into different schools and have ultimately shaped the history of the discipline. While a long tradition of thought has contributed to developing economic models of individual behavior, dissenting groups have recurrently pointed to its neglect of other important motives of human action; while most economists since a very early stage have promoted a conception of the market as a self-adjusting social mechanism capable of coordinating individual actions at best, critics have often raised doubts on its merits; and while a majority has often supported pro–free market arguments against government intervention, the opposite position has sometimes prevailed. In outlining these developments, similarities and differences between past and present theories will be emphasized whenever possible, with the help of HET literature and in an effort to stress the insight that may come from both of the above-outlined approaches.

Further readings include classic works such as Robert Heilbroner's (1999) *The Worldly Philosophers,* Joseph Schumpeter's (1954) *History of Economic Analysis,* and Mark Blaug's (1997) *Economic Theory in Retrospect,* together with recent reference books such as *A Companion to the History of Economic Thought* (Samuels, Biddle, & Davis, 2006) and *The History of Economic Thought: A Reader* (Medema & Samuels, 2003). The main scholarly journals in the field are *History of Political Economy, European Journal of the History of Economic Thought,* and *Journal of the History of Economic Thought.* The Web site of the History of Economics Society (http://historyofeconomics.org) and the HET page of the New School for Social Research (http://cepa.newschool.edu/het) also offer information. Finally, Liberty Fund has republished at affordable prices many of the great books that have made the history of the discipline, providing some of them online at its Library of Economics and Liberty (http://www.econlib.org).

## HET Through 1925

Scholars in Antiquity and the Middle Ages thought a great deal about trade, money, prices, and interest rates, but an autonomous discipline developed only toward the late seventeenth to early eighteenth centuries. The early designation of *political economy,* proposed by Antoine de Montchrestien in 1615, was later replaced by *economics* and refers today to a specific subfield only, but it has the merit of stressing a persisting feature of the discipline as a whole—namely, its linkages with public policy and the role of the economist as an adviser to the policy maker. This specificity still matters and distinguishes economics from other social sciences.

Despite the interest of the early literature (see, e.g., Hutchison, 1988), a detailed account of it would be beyond the scope of this chapter, and the more traditional convention of starting from the late eighteenth century will be followed. Focus will be on Adam Smith (1723–1790), who is widely regarded as one founder of the discipline; the remainder of this section will outline the development of economic thought in the nineteenth and early twentieth centuries, with the emergence of the so-called classical and then the neoclassical schools.

### Adam Smith

In his 1776 *Wealth of Nations,* Smith laid the foundations of what would become basic principles of economists' understanding of individual behavior, the market mechanism, and the role of markets vis-à-vis governments. Smith was the first to explicitly characterize individual economic behavior as *self-interested behavior,* admitting that it is people's desire for a gain that explains work, production, and ultimately the existence of an economic system:

> It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages. (Smith, 1776/1981, pp. 26–27)

Self-interest was initially thought to be at odds with another principle that Smith had developed in his *Theory of Moral Sentiments* (1759/1982)—namely, sympathy between

human beings who, by putting themselves imaginatively in the place of others, understand their feelings and expectations and are moved to act accordingly (e.g., to give to those who are in need). This apparent contradiction, known in the literature as *Das Adam Smith Problem,* has been largely resolved by recent scholarship that has rather stressed how self-interest and sympathy emphasize different aspects of human nature, whose relative importance varies depending on the situation. They constitute two instances of a unique framework for thinking about human behavior, in which the individually centered, self-interested component is accompanied by an interpersonal dimension, so that it becomes possible to account for various forms of behavior, from trade and profit-seeking actions to philanthropy. Reconciliation of these two aspects of Smith's thought makes him the father of economics in a broad, comprehensive sense: Although the discipline was long viewed as the systematic analysis of the behavior of self-interested individuals, today's research (especially in behavioral economics) tends to integrate forms of prosocial behavior into economic analysis.

Smith's work also contributed to shaping economists' view of the market as a coordinating device in a world in which private property and freedom enable individuals to make self-interested decisions autonomously, without ex ante coordination by some outside (political) authority: The market is a social mechanism that ensures, ex post, that individual decisions are consistent with one another and generate an orderly result. Smith's "invisible hand" metaphor has often been recognized as an effective representation of this mechanism:

> by directing [ . . . ] industry in such a manner as its produce may be of the greatest value, he [man] intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it. (Smith, 1776/1981, p. 456)

Because individuals are not isolated but part of a larger human community, their actions have unexpected or *unintended consequences* at the system level. Individuals take into account only their self-interest, yet their choices affect others and trigger a chain of interactions that eventually affect society as a whole, well beyond their original intentions. Strikingly enough, Smith argues that this spontaneous process does not lead to chaos but to harmony: Self-interest may not seem a noble motivation, yet it triggers consequences that benefit society even more than those arising from benevolence. Thus, there is no need for a strong state power that would impose social order from above, as argued by Thomas Hobbes (1651/1996). The idea of unintended consequences and the possible reconciliation of individual self-interest and social good, first articulated by Smith, have been at the core of subsequent economic reflection—which is another reason why Smith is credited as a founder of the discipline.

While acknowledging the merits of the market, Smith did not deny the need for a solid government. In particular, he insisted that government should ensure the basic conditions that allow markets to function properly, primarily protection of private property and enforcement of contracts (Smith, 1776/1981, p. 910). The government should also be in charge of surveillance against what we would call today unfair competition and other abuses (Smith, 1776/1981, p. 145).

## The Smithian Heritage and the "Classical" School

Smith's seminal work stimulated much reflection. On individual behavior, the idea that individuals act to satisfy their self-interest gradually developed into the individual optimization principle that is at the basis of today's textbook microeconomics (see below). This principle has often been criticized as restrictive, not taking into account the multifaceted motivations that drive human behavior. Yet historians of economics have highlighted that in the nineteenth century, the economic model of individual behavior had the merit of supporting an egalitarian perspective that had great impact on political debates. If the same scheme holds for all individuals, they are all equal and have the *same capacity for decision making;* observed differences, if any, depend only on incentives, chance, and history. This egalitarian view was shared by most economists of the time and strongly contrasted the (then also widespread) hierarchical stance that regarded the lower classes of society and supposedly "inferior" ethnic/racial groups as less capable of making decisions and thus in need of guidance (Peart & Levy, 2005). Interestingly, the infamous "dismal science" designation of political economy was originally an accusation against economists' antislavery orientation, which resulted from their belief that all humans are equal (Levy, 2001).

On the coordinating mechanism, reflection was motivated by the questions that Smith had left open as well as by the socioeconomic transformation brought by the Industrial Revolution. David Ricardo (1772–1823) provided one of the finest analyses of the relationship between prices and quantities; although he did not use mathematics, his compelling arguments raised the level of rigor in the subject and set basic standards for later modeling. His *On the Principles of Political Economy and Taxation* (1817/2004) offers an in-depth analysis of the effects of scarcity on price formation. Suppose some quantity of produce (say, a ton of corn) can be obtained with a given amount of labor and other inputs (seeds, water, fertilizers) on fertile land. Production of a ton of corn on less productive terrain requires larger amounts of inputs, so production costs are higher. Hence, self-interested producers will exploit fertile lands first and will extend production to lower quality lands only when demand is so strong that there are no spare high-quality lands to satisfy it. In such cases, high consumer demand

pushes up the market price of corn until it covers production costs on the worst fields; consequently, price exceeds the cost of producing corn on the best fields and earns a "rent" for their landlords. While this description is highly simplified and does not take into account potentially relevant factors such as technological progress, it still allows applications to present-day natural resource economics. An example is oil, whose extraction costs differ in different areas, so the worst oil fields are profitable only when global demand is strong and prices rise.

Prices and quantities also vary depending on the competitive conditions that prevail in a market. It was already known that in monopoly situations, prices are higher and quantities are lower than in cases in which several firms compete, but no rigorous explanation of this phenomenon was available until Augustin Cournot (1801–1877) with his *Researches Into the Mathematical Principles of the Theory of Wealth* (1838/1927). A pioneer of the use of mathematics to guide economic reasoning in times when few did so, he illustrated how the sole seller of a good is able to control the entire market and hence extract a monopoly rent. In the case of two or more sellers, Cournot highlighted the importance of strategic interactions, so that each seller's decision depends on other sellers. The market is in equilibrium when each firm's output maximizes its profits given the output of other firms—a notion of equilibrium that has been acknowledged to prefigure that of (pure-strategy) Nash equilibrium in game theory. The author also thought that the weight of strategic interactions declines as the number of sellers increases, so when many of them compete, no one is capable of exerting any influence on market prices. Cournot was the first to prove that under competitive conditions, the quantity produced is such that its market price equals (what we would call today) its marginal cost.

Economists of this time period are often referred to as "classical" economists. Other prominent classical writers include Thomas Malthus (1766–1834) and John Stuart Mill (1806–1873). There have been controversies on the definition of the classical school, its timeframe, and scope: Smith and Ricardo are considered among its main contributors, while Cournot is less frequently included, although there are similarities and differences between all three. On the whole, these writers were mainly interested in production/supply forces, working conditions, and the relationship between wages and profits, while they placed relatively less emphasis on utility, consumption, and demand. Although Ricardo's rent theory relied on demand to determine whether less productive resources could be profitably used, and Cournot went as far as to draw supply-and-demand diagrams, these writers did not derive demand from utility; even Cournot with all his mathematics regarded it only as an aggregate relationship calculable from expenditure data. Arguably, it is not that their arguments were underdeveloped but that the classical school primarily pointed to the influence of the whole economic system on individual behavior, rather than the other way round, and focused on the differential impact of conditions of production on individuals according to their position as workers, capitalists, or landowners. This viewpoint had the merit of calling attention to problems related to income distribution and the possible tension between wages and profits.

Early nineteenth-century policy debates focused on free-market principles and the role of the state in countering potential negative effects of markets. One key controversy concerned unemployment, of which external trade was one perceived cause, to the extent that increased low-cost imports might have resulted in national workers being displaced by cheaper foreign workers; similarly, technological innovation could be conducive to displacement of workers by machines. Would the market mechanism self-adjust to reabsorb the workers left idle by an opening of the country to external trade and/or a technological shock? These questions are important for the well-being of a country and, since then, have recurred several times in the history of economics. A prominent contributor was Ricardo, who first claimed that market adjustments would be sufficient and then admitted that consequences for the working classes were likely to be hard, at least for some time—although he still believed that restraining technical progress and free trade would have been detrimental to the country.

Like Ricardo, many supported free trade despite its possible inconveniencies and the need for government to intervene in some cases. However, this time period also saw the development of socialist ideas in reaction to the conditions of workers in the new industrial age and the classical economic thought that accompanied it. Influenced by classical authors but at the same time critical of them, Karl Marx (1818–1883) highlighted the internal contradictions of the current social relationships of production, the conflict between labor and capital, and the historical tendencies that brought about the modern economic system but also generated tensions that eventually led to its collapse. Marx's *Capital* (1867) attracted many followers in economics and also inspired political action directed at radical social, economic, and political change.

## The Late Nineteenth Century and the "Neoclassical" (Marginalist) School

From the second half of the nineteenth century onward, increased emphasis was put on consumption rather than production only, with the introduction of a notion of utility as a measure of individual satisfaction from the consumption of goods or services. Some reflections on utility had already appeared, with the idea that the problem of political economy and the ultimate purpose of all productive activities is to satisfy human wants at best. Yet it was long before utility could be fully integrated within economic models, not least because at first glance, it may have appeared as a subjective, qualitative notion devoid of any objective, let alone quantifiable, attribute. The solution came from reinterpreting utility not as an absolute but as a *relative* magnitude, varying from one individual to another and for each

individual, depending on the available quantity of a good. One could thus distinguish the total amount of utility from "marginal" utility—namely, the change in the level of utility that results from a given increase in the quantity of the good. Marginal utility was thought to diminish with the quantity consumed, reflecting the capacity of individuals to order the possible uses of successively acquired units: For instance, one would reserve the first gallons of water for drinking and the successive ones for personal hygiene, for housekeeping, and finally for watering plants. In passing, this assumption solved what earlier thinkers considered a paradox—the fact that useful goods such as water or air have low market value: The reason is their abundance, which means that the last increment in quantity generates an extremely small increase in utility. These results suggested an interpretation of self-interested behavior in terms of attempts to raise one's utility to its highest possible level and were obtained independently, in 1871–1874, by William S. Jevons (1835–1882) in Britain, Carl Menger (1840–1921) in Austria, and Léon Walras (1834–1910) in Switzerland.

The importance of thinking in terms of marginal variations rather than total magnitudes proved so useful to account for utility and demand that it was subsequently extended to supply. In fact, notions of marginal productivity and marginal cost of production, as opposed to total productivity/cost, had already been introduced (e.g., by Cournot) but were refined and generalized in the 1890s by, among others, John B. Clark (1847–1938), Philip H. Wicksteed (1844–1927), and Knut Wicksell (1851–1926). Marginal reasoning seemed so important that the economic thought of this time period is often referred to as *marginalism.*

Accounts of the market mechanism of this time period place emphasis on the *symmetry* of supply-and-demand factors and on the resulting equilibrium. Individual demand and supply are derived, respectively, from agents' calculations of utility (for consumers) and profit/cost (for firms), and market supply and demand are obtained by aggregating all individual values. When market supply equals demand, the market is in equilibrium—that is, the decisions of all households and all firms are consistent with one another. These common traits can be combined with different assumptions to give rise to various models of the market. Alfred Marshall (1842–1924) is renowned for developing a "partial equilibrium" approach, focusing on the study of a single competitive market and illustrated with the help of price-quantity diagrams in which demand decreases and supply increases with price. The intersection of the supply-and-demand schedules identifies equilibrium—a price at which supply equals demand and the market clears (Marshall, 1920). The partial equilibrium approach provides a tractable framework to study the relationship between price and quantity; however, it is based on the restrictive assumption that changes in the price of a good have repercussions on the quantity of that good only, ruling out the possibility that a variation in the price of a good will have an impact on the demand and supply of substitutes and/or complements. Hence, it can be taken at most as an approximation, not as a rigorous analytical device. In contrast, interdependencies among markets were a key concern for Walras (1874/1977), who tried to model agents who allocate their budgets to the purchase of multiple goods so that changes in the market price and/or quantity of one good are likely to have repercussions on the markets for other goods. His notion of "general equilibrium" is directly derived from this view and corresponds to a situation in which supply equals demand *on each market,* so that all clear simultaneously.

A closely related, though distinct, question is whether and how actual trade practices will drive prices and quantities toward equilibrium. Again, Marshall and Walras provided different answers. Walras (1874/1977) proposed a model of auctions in which at given prices, all traders declare the quantity of each good that they wish to buy or sell at those prices; if with these quantities, supply equals demand on each market, then this is the general equilibrium and trade takes place; if not, prices are adjusted in such a way that they diminish where supply exceeds demand and increase in the opposite case; at the new set of prices, traders announce again the quantities that they wish to buy or sell, and the process (which he labeled *tâtonnement,* a term still used in the literature) starts again, until equilibrium is reached. In short, transactions take place simultaneously, at equilibrium only, so that the same prices apply to all traders. Instead, Marshall (1920) had in mind a sequence of bilateral transactions on a single market, in which each pair of traders negotiates a price and each transaction withdraws some units from the market so that lesser quantities are available for later trades—in other words, the conditions under which traders negotiate are altered at every step. Such changes gradually dampen price adjustments until they reach the level that corresponds to the intersection of supply and demand. Here, transactions occur sequentially, in disequilibrium, at prices that may differ from one pair of traders to the other.

Is there continuity or rupture between the classical school of thought and the marginalist—also known as "neoclassical"? The emergence of the latter current of thought used to be referred to in HET as a "revolution," thus suggesting a major change, which some argue is primarily due to the postulated symmetry between supply-and-demand conditions rather than to the use of marginal concepts, yet important features of neoclassical thought and elements of reflection on utility and demand were anticipated by earlier authors. Today's HET scholars mostly believe that there was no such thing as a sudden transformation of the discipline but a long, slow transition; key marginalist concepts appeared early, but it took long before they were systematized into a coherent, comprehensive framework. In turn, neoclassical economics does not constitute a single theory but rather a family of approaches: The market models of Marshall and Walras are examples of such differences.

Openness to more rigorous thinking and increased use of mathematics have been often thought to characterize neoclassical theories; an indication of this tendency is the renaming of the discipline in the late nineteenth century from *political economy* to *economics,* primarily at the initiative of Marshall. However, qualifications should be introduced to the extent that some earlier writers such as Cournot had already used some mathematical tools in their analysis (Theocharis, 1993); conversely, late nineteenth-century economists were not unanimous on the desirability of using mathematics, and Marshall himself limited his quantitative expressions to a minimum. It took long before the use of mathematics became standard in the profession, and debates on the legitimacy of using mathematical tools in the study of human behavior and society have been recurrent since then.

## HET Since 1925

This section outlines the development of neoclassical economics in the twentieth century, with focus on its two hallmarks of individual optimization and market equilibrium. The section also includes an overview of the emergence of macroeconomics and how it gradually came to incorporate the principles of optimization and equilibrium.

### Neoclassical Economics: Individual Optimization and Market Equilibrium

The neoclassical concept of utility was refined over time, and mathematical models of utility maximization under a budget constraint saw the light. Progressively, the constrained maximization model was extended to the study of all individual decision units, on both the demand and the supply side of the market, and became the basis of all analyses of individual economic behavior. It gradually came to be understood as rational behavior—choosing the best possible means to achieve one's ends. This opened the way to a conception of economics based on two pillars: optimizing behavior of agents (both consumers and firms with, respectively, utility and profit as objective functions) and equilibrium of markets. The contribution of Paul Samuelson (1915–2009) was essential to these developments and mainly consisted in rewriting many problems of economics as maximization problems, with extensive use of mathematics. From the 1940s onward, Samuelson's effort to show that apparently diverse subjects have the same underlying structure and can be treated with the same mathematical tools gave unprecedented unity and coherence to the discipline.

The optimization model has not been beyond dispute, though: At least since the 1950s, Herbert Simon (1916–2001) and others contended that actual decision makers lack the cognitive capacities to solve maximization problems and rather content themselves with "satisficing" behavior, choosing options that are not optimal but make them happy enough. Along these lines, they developed a "bounded rationality" approach as an alternative to the seemingly strong rationality requirements of the individual maximization model.

Regarding market models, Marshall's partial equilibrium approach continued to be used in applied economic studies, but it was the general equilibrium model that most attracted the attention of economic theorists during this time period. Its extraordinary development after World War II was largely due to the introduction into economics of highly advanced mathematical tools and of a new way of thinking about mathematics (Weintraub, 2002a). The new tools allowed for a sophisticated refinement of Walras's approach, named the Arrow-Debreu-McKenzie model after its main contributors. A major achievement in the 1950s was a formal proof of existence of equilibrium. The difficulty was that it was not enough to show that the system of simultaneous equations representing equality between supply and demand in all markets has a solution: For this solution to be meaningful economically and not only mathematically, it was also necessary to prove that equilibrium prices and quantities are nonnegative. By demonstrating that it is indeed the case, it was established that the notion of a set of prices that clear all markets is consistent (i.e., that the notion of equilibrium of a system of interrelated competitive markets is not void).

Another success for general equilibrium theory was the mathematical proof of the so-called two theorems of welfare economics—a modern reinterpretation of Smith's "invisible hand." The first theorem states that a general equilibrium corresponds to a socially optimal allocation of resources, and the second states that, under some conditions, any socially optimal allocation of resources can be sustainable by a general equilibrium. These results shaped policy discussions for long: They amounted to rigorously establishing the properties of the free-market mechanism that earlier economists had put forward intuitively—namely, the idea that a market-based solution to the problem of allocating scarce resources produces a desirable outcome for all and cannot be superseded by any other alternative. The two theorems made a strong case for the free market but also enabled clear identification of cases where government intervention is legitimate: Whenever the assumptions that support the two theorems (e.g., competitive conditions) are not met, the market may fail to yield efficient outcomes ("market failures"), and the government should step in.

In the 1950s and 1960s, such progresses put the Arrow-Debreu-McKenzie model at the center of the stage and increased confidence in its potential to provide the whole of economics with rigorous mathematical foundations. However, problems started with attempts at proving two other key properties of equilibrium—namely, stability and uniqueness. The question of stability was meant to ensure that after an exogenous shock, the market mechanism is capable of generating endogenous forces that bring it back to equilibrium; if equilibrium exists but the market cannot find it, then arguments for free markets are harder to make. In addition, if uniqueness is not guaranteed, it is unclear where an adjustment process might drive the system after a

shock; besides, some equilibriums may be unstable. It became soon clear, though, that formal proofs of stability and uniqueness could be obtained only under very restrictive, unrealistic assumptions. Critics stressed that these results reveal that with all its mathematical underpinnings, general equilibrium theory did not truly succeed in improving knowledge of how the market mechanism works and how prices adjust in response to variations of supply-and-demand conditions (Ingrao & Israel, 1990). Today the theory is still part of economists' education, but research in this field has entered a phase of relative decline.

## Keynes and the Emergence of Macroeconomics

In the aftermath of the Great Depression, macroeconomics also entered the scene. John Maynard Keynes (1883–1946), one of the fathers of the new approach, is among the most influential economists of the twentieth century. His *General Theory of Employment, Interest and Money* (1936/1997) proposed a new way to look at economic systems, one that placed emphasis on quantity adjustments at given prices, rather than on the supposed capacity of price adjustments to equilibrate markets, and focused on aggregate relationships rather than on individual behavior. The new approach was rooted in the belief that the macrolevel of analysis differs in nature from the microlevel and contented itself with the use of simple behavioral assumptions that do not require optimization: Consumers spend a fraction of their income and save the remaining part, and firms' investment decisions depend inversely on the interest rate. In this way, it hoped to tackle the question of unemployment that was crucial at the time. The neoclassical conception was ill-suited to explain why some people could be involuntarily unemployed for long: It regarded labor as hardly different from any other good, so that market price adjustments should in principle bring the system back to its full-employment equilibrium after a temporary shock. In contrast, Keynes stressed the specificity of labor relative to other goods and suggested that the level of employment may depend less on prices than on aggregate demand (i.e., the total expenses of an economic system). In this perspective, an economy may be unable to deliver full employment, even if all markets for goods clear in the long run: Underemployment may be its normal state. In this sense, the book challenged the idea that markets are capable of self-regulation and built a theoretical framework that legitimated increased government intervention to stimulate the economy. Policy measures could take various forms, ranging from increased public spending to lower interest rates to encourage investments and thus raise demand for labor.

The book had enormous success and changed the way economists and policy makers looked at the role of governments in a market economy: Not only did it inspire a significant amount of research work, but it was also at the basis of economic policies in Western countries after World War II, so it is sometimes referred to as a Keynesian "revolution." Policies of demand stimulation along Keynesian lines were enhanced by the parallel development of macroeconometric techniques that made it possible to assess the state of an economic system and to estimate the impact of government interventions.

However, the neoclassical approach was not completely abandoned, and in particular, a group of economists tried to reconcile it with Keynes's view. A hallmark of this tendency is the model known as IS/LM, designed in 1937 by John Hicks (1904–1989) to represent key principles of the *General Theory* in the form of a system of simultaneous equations reminiscent of those of general equilibrium theory. Indeed, the model succeeded in capturing some major aspects of Keynes's thought, but at the same time, its partly neoclassical roots made it less suited to account for the possible existence of unemployment in an economy that is otherwise in equilibrium. Later, the IS/LM model was enlarged to take into account international transactions (the so-called Mundell-Fleming model) and was accompanied by a Phillips curve that explained the behavior of prices on the basis of an inverse relationship between inflation and the level of employment. Starting in the mid-1950s, a consensus emerged around this approach at least in the United States, where it constituted a basis for descriptive economic analysis and for policy advice. It coexisted with neoclassical microeconomics and became known as the *neoclassical synthesis,* a designation commonly attributed to Samuelson.

Keynesianism came under attack after the 1973 oil crisis and the ensuing nasty combination of inflation and unemployment for which it seemed to have no remedies: Demand policies would reduce unemployment but would lead to higher inflation, while public expense cuts would tame inflation but would raise unemployment. This period saw the rise of an alternative approach, known as *monetarism* and primarily associated with Milton Friedman (1912–2006). Monetarism revived the pre-Keynesian belief that market economies can regulate themselves without any need for government intervention and brought to light a strong relationship between money creation and inflation, so that an economy may be destabilized if the authorities print too much money: It followed that the focus of economic policies should be solely on keeping the quantity of money under control and that active demand policies are useless, if not in fact damaging. Monetarism spread widely in the early 1980s and had a strong influence on policy making. Keynesian ideas did not completely vacate the scene, though: Many became convinced that government policies can still have a temporary effect and that the Keynesian framework of analysis holds in the short run, while the monetarist framework holds in the long run.

This compromise was challenged by Robert Lucas (born 1937), who made a strong case for unifying the foundations of economic theory through an extension to macroeconomics of the microeconomic assumptions of the rational behavior of individuals and of the self-equilibrating capacity of markets. Agents make optimal choices: In particular, they form expectations about the state of the economy by taking into account all available information and by processing it in the best possible way, so that they can be

called "rational expectations." More precisely, agents make consumption or investment decisions that take into account the model of the economy as well as government policies, so that they reach equilibrium immediately and their expectations are validated. If all agents behave in this way, the economy is always in equilibrium. A major implication of this view is that economic policies are ineffective even in the short run because they are anticipated by agents and are accounted for in their decisions. Only unexpected policies that take individuals by surprise can move the economy from one state to the other—but this means that to be effective, policies must be occasional and unsystematic or else they will be detected, so the scope for governments to steer the economy becomes extremely limited. A consequence of this view is the invalidity of macroeconometric models that purported to evaluate the effects of public policies with the help of aggregate data: If agents take into account policies in their decision making, their behavior is not policy invariant so that existing observations may not predict future choices well. Only a sophisticated model of how individuals make optimal decisions based on their expectations can offer reliable predictions.

The rational expectations school of thought tried to give greater coherence to the discipline by basing both micro- and macrotheories on the two main pillars of individual optimization and market equilibrium. This choice responded to a widespread demand for more rigorous economic theorizing but also reflected renewed confidence in the functioning of the free-market mechanism and skepticism with regard to government intervention; in this sense, it represents a comeback for pre-Keynesian attitudes. Since then, most developments of macroeconomics have reflected this tendency—with the development, among other things, of the real business cycles approach by Finn E. Kydland and Edward Prescott in the 1980s.

Although the great majority of macroeconomists have now recognized the need to firmly ground macroeconomic theorizing on sound microeconomic foundations, many disagree with the pro–free market orientation of these currents and have tried to develop alternative approaches that would still be based on rigorous microfoundations but would lead to Keynesian results, most prominently by showing the possibility for unemployment to persist in an equilibrium economy. These approaches, commonly referred to as *New Keynesian,* have brought to light characteristics of the economy that might lead to this result, ranging from implicit contracts, efficiency wages, and coordination failures to imperfect competition. An overview and an appreciation of their contributions are provided in De Vroey (2004).

Over time, a consensus gradually has been established around general equilibrium theory and macroeconomics with rigorous microfoundations, which have come to be identified as the core of the discipline—what some now call "mainstream" economics. They are now at the basis of economics education and constitute a reference for the profession as a whole. Since their introduction, training in these fields has contributed to raise the level of rigor in economics reasoning and to spread the use of mathematical and quantitative tools.

## History of Recent Economics

Research in mainstream economics is still active, even if there has been a relative decline in recent years. This has paralleled a tendency to increasing diversification of approaches, methods, and topics, which has seemingly reversed the twentieth-century trend toward unification of the different parts of the discipline. Still, economists tend to have in common an enduring emphasis on mathematical tools and formal reasoning.

A detailed account of the different emerging approaches to economics would be beyond the scope of this brief account of HET, but more information can be found in other chapters in this handbook. This section instead will provide an overview of some significant developments and how they are challenging established knowledge in economics.

Models of rational behavior have put the accent on the *strategic dimension* of rationality in situations where agents make decisions whose success depends on the choices of others. This shift in emphasis results from the rise of game theory as a challenger to established microeconomics, which has been spectacular in recent years even though the origins of the theory date back to (at least) the 1940s. Applications of game theory include bargaining, imperfect competition, and questions at the interface between economics and other sciences, such as social network formation, the emergence of social norms, and voting systems.

Assumptions of individual rationality do not go unquestioned, though. The stream of research that is known as "behavioral economics" has provided substantial evidence that humans often violate some implications of optimization models and has tried to develop more realistic psychological approaches to the study of individual behavior. Some researchers, in particular, have focused on how happiness and individual satisfaction, as well as prosocial and cooperative attitudes, may be important determinants of individual behavior that were not fully accounted for in older maximization models. To do so, new sources of information have been exploited, notably experimentation and analysis of survey data with the help of increasingly sophisticated microeconometric techniques. While these fields remained marginal for a while, they are now recognized parts of the discipline and attract an increasing number of young economists.

Models of the market have been greatly enriched by a detailed study of auctions and other mechanisms of allocating goods. Part of the motivation for these studies in the 1990s and early 2000s was the need to design trading mechanisms that would help governments to privatize companies, infrastructures, and other facilities that they

previously owned. To some extent, economists' work in this area resembles that of engineers at the service of the government—a new role that, nevertheless, renews the time-honored image of the political economist as an adviser to the policy maker. These studies depart from the general equilibrium tradition in a double sense: First, they highlight the importance of trading institutions to yield socially desirable outcomes, instead of abstracting them away, and second, they signal a tendency to focus less on interdependencies and rather concentrate on single markets—what Marshall modeled in a "partial equilibrium" perspective.

At the macrolevel, greater emphasis has been placed on economic governance. The conditions under which governments can ensure protection of property rights and enforcement of contracts, already emphasized by Smith as key requirements for market economies to function properly, have been studied in greater depth from the 1990s onward. Focus on governance and institutions sometimes accompanies criticisms of pro–free market principles but sometimes supports the free-market tradition of thought by providing a more precise definition of how the government can create conditions for markets to function properly.

## Conclusion

To conclude, it can be said that the discipline advances over time with the progressive introduction of new tools, new approaches, and an improved understanding of key concepts. Yet some questions are recurrent and constitute some of the great, unresolved dilemmas of contemporary society. This chapter has emphasized the problems of individual economic behavior, the functioning of the market mechanisms, and the place of the market vis-à-vis the government. The answers provided at different epochs, though based on different arguments and different sources of evidence, often have elements in common—partly because these are issues that have major philosophical and political implications. By accounting for the circumstances in which a variety of responses have emerged in the past, HET can contribute to today's reflection on these issues. As Keynes (1936) once wrote,

> The ideas of economists and political philosophers, both when they are right and when they are wrong, are more powerful than is commonly understood. Indeed, the world is ruled by little else. Practical men, who believe themselves to be quite exempt from any intellectual influences, are usually the slaves of some defunct economist. (p. 383)

## Notes

1. For an overview of methodological debates in HET, readers may wish to consult Weintraub (2002b).

## References and Further Readings


Blaug, M. (1997). *Economic theory in retrospect* (5th ed.). Cambridge, UK: Cambridge University Press.

Cournot, A. A. (1927). *Researches into the mathematical principles of the theory of wealth* (N. T. Bacon, Trans.). New York: Macmillan. (Original work published 1838)

De Vroey, M. (2004). *Involuntary unemployment: The elusive quest for a theory.* London: Routledge.

Heilbroner, R. (1999). *The worldly philosophers* (Rev. 7th ed.). New York: Simon & Schuster.

Hobbes, T. (1996). *Leviathan or the matter, form, and power of a commonwealth ecclesiastical and civil* (Rev. student ed.). Cambridge, UK: Cambridge University Press. (Original work published 1651)

Hutchison, T. W. (1988). *Before Adam Smith: The emergence of political economy 1662–1776.* Oxford, UK: Blackwell.

Ingrao, B., & Israel, G. (1990). *The invisible hand: Economic equilibrium in the history of science.* Cambridge: MIT Press.

Keynes, J. M. (1936). *The general theory of employment, interest and money.* London: Macmillan.

Levy, D. M. (2001). *How the dismal science got its name: Classical economics and the Ur-text of racial politics.* Ann Arbor: University of Michigan Press.

Marshall, A. (1920). *Principles of economics* (8th ed.). London: Macmillan. Available at http://www.econlib.org/library/Marshall/marP.html

Marx, K. (1867). *Capital: A critique of political economy.* English translation based on 4th ed. Available at http://www.econlib.org/library/YPDBooks/Marx/mrxCpACover.html

Medema, S., & Samuels, W. J. (2003). *The history of economic thought: A reader.* London: Routledge.

Peart, S. J., & Levy, D. M. (2005). *The "vanity of the philosopher": From equality to hierarchy in post-classical economics.* Ann Arbor: University of Michigan Press.

Ricardo, D. (2004). *On the principles of political economy and taxation* (P. Sraffa & M. Dobb, Eds.). Indianapolis, IN: Liberty Fund. (Original work published 1817)

Samuels, W. J., Biddle, J. E., & Davis, J. B. (Eds.). (2006). *A companion to the history of economic thought.* Malden, MA: Blackwell.

Schumpeter, J. A. (1954). *History of economic analysis* (E. Boody Schumpeter, Ed.). New York: Oxford University Press.

Smith, A. (1981). *An inquiry into the nature and causes of the wealth of nations* (2 vols., R. H. Campbell & A. S. Skinner, Eds.). Indianapolis, IN: Liberty Fund. (Original work published 1776)

Smith, A. (1982). *The theory of moral sentiments* (D. D. Raphael & A. L. Macfie, Eds.). Indianapolis, IN: Liberty Fund. (Original work published 1759)

Theocharis, R. D. (1993). *The development of mathematical economics: The years of transition, from Cournot to Jevons.* Houndmills, UK: Macmillan.

Walras, L. (1977). *Elements of pure economics: or the theory of social wealth* (W. Jaffé, Trans.). Fairfield, NJ: A. M. Kelley. (Original work published 1874)

Weintraub, E. R. (2002a). *How economics became a mathematical science.* Durham, NC: Duke University Press.

Weintraub, E. R. (Ed.). (2002b). *The future of the history of economics.* Durham, NC: Duke University Press.

# 2

# Economic History

Sue Headlee
*American University*

Economic history is the series of social arrangements and physical processes by which human societies produced the material conditions of human life since the emergence of the human species. The discipline of economic history is the study of this series of arrangements and processes, although much of the discipline is devoted to the study of the development of modern economic growth. The reason for this is that modern economic growth brought with it sustained and accumulating increases in the per capita wealth of human societies. Before modern economic growth, any improvement in productivity led to an increase in the population, not an increase in the standard of living.

In this chapter, human economic history will be examined through the lens of the discipline of economic history. First, theory is analyzed, and then empirical evidence. The theory and evidence sections are followed by policy implications, future directions for research, and a conclusion.

## Theory

Neoclassical economic theory was largely developed in the nineteenth and twentieth centuries at the time of industrialization of the West. It is able to explain the increase of total output and the output per capita for societies with market economies experiencing modern economic growth. Simon Kuznets (1968) defined modern economic growth as sustained and faster growth of output per capita than the rate of growth in earlier periods of history. Neoclassical economic theory does not explain long-term growth or growth of societies where the market is not the predominant mechanism to allocate resources.

Economic growth is caused by an increase in the amount or quality of capital or labor used in production, an increase in the ratio of capital to labor, and technological innovation, according to Robert Solow's (1970) theory of economic growth. Douglass North (1981) added a theory of institutions to the theory to make it possible to analyze longer term growth, starting before markets and aiming to understand how societies came to have markets allocate resources. North thus theorized about how humans advanced from hunting and gathering to the discovery of agriculture and the subsequent rise of ancient civilizations such as ancient Greece and Rome, the rise and fall of feudalism in Europe, and finally the era of early modern Europe. With a theory of institutions, North could explain the distribution of the costs and benefits of economic growth.

As a field of economics, the discipline of economic history has focused on the causes and effects of modern economic growth in the West. This is important because in fact modern economic growth arose in the West and then transformed the world, creating the industrial civilization that we live in today. What follows is a survey of the factors most widely thought to cause modern economic growth.

### Causes of Modern Economic Growth

*Expansion of Markets and Trade*

Adam Smith (1776/1976) wrote that the greater the extent of the market and the greater the development of the division of labor and specialization, the greater the wealth of nations. Since Smith, neoclassical economists have focused on markets, the market system, and the price mechanism as the foundations of modern economic growth.

Neoclassical economic theory states that when output and inputs to production are allocated by markets efficiently, this stimulates growth. In addition, relative prices guide economic agents to make production and distribution decisions efficiently. David Ricardo (1819) put forth the comparative advantage theory of why countries trade. This theory held that it was to the advantage of all countries to trade by specializing in the production of a good in which they had a comparative, even if not absolute, advantage. Increasing international trade causes economic growth.

*Evolution of Institutions*

North won the Nobel Prize in Economics for adding the causal factor of the evolution of institutions to the neoclassical growth model. A key institution is the regime of property rights, which gives economic agents incentive to save and invest in physical, financial, and human capital. Political institutions evolve that specify and enforce these property rights. Institutions, like markets, provide incentives. Economists believe that when there are positive incentives to do so, people will respond with behavior that leads to growth. If property rights are secure, farmers will invest in improving their farms.

*Rise of the Modern State*

Among the institutions conducive to modern economic growth, of great importance is the modern state. Premodern states are a drag on progress because they favor the elite, aristocrats, and landowners, giving them rent (income from power and owning land) rather than gains from investment in production. Modern states, more representative of the middle classes, can implement economic policy to favor domestic markets. This leads to an increase in the standard of living of the population. Government policies that help economic growth include the investment in infrastructure, such as railroads and ports, and in human capital, such as literacy training for adults and the spread of primary education for children.

*Accelerated Technological Change*

Thomas Malthus (1798/1993) held that incomes do not rise when there is an increase in productivity. He argued that populations tend to grow beyond the capacity of resources and that societies use an increase in production to support more people, rather than to increase the standard of living. Historically, much of the world has stayed in a static economic condition or has gone through waves of expansion followed by decline because of this Malthusian trap. Malthus lived before the great acceleration of technology that made modern economic growth possible.

Joel Mokyr (1990, 2002) has written that the important characteristic of the Industrial Revolution was its accelerating and unprecedented technological change. Technological innovation increases the productivity of labor, making an increase in income per capita possible. Technological change has developed slowly throughout history, but in the late eighteenth century, there was a dramatic speedup in the pace and the ability to sustain it. The Industrial Revolution was a pivotal event in human history. Technological change revolutionized manufacturing, agriculture, and transportation in the nineteenth and twentieth centuries.

*Human Capital and the Stock of Knowledge*

Gary Becker (1975) stressed the role of human capital in causing economic growth. Human capital is what economists call the result of investing in education and training, just as investing in productive equipment is called capital formation. Solow's (1970) growth model shows that this investing in education improves the quality of the labor input and thus adds to the productive capacity of the economy. Mokyr (1990, 2002) provides a history of the development of the stock of knowledge in the West that enabled the elites of societies to advance the mastery of nature for the benefits of the masses of the population.

*Demographic Changes*

Becker (1975) found that under certain circumstances, people reduce the number of children they have. They may do this because they no longer need the labor of children on the farm or for security in old age. They may do it to have a higher standard of living for themselves and/or to have "higher" quality children. This makes it possible for the economy's output to grow faster than the population. There is a feedback loop here in that as income per capita grows, people reduce family size further. In some societies, the demographic transition from women having many children to having fewer children comes before economic growth, and in some it comes after that growth. North (1981) speaks of this process as lining up individual and social costs and benefits of having children.

*Evolution of Capitalism and Industrialization*

Under capitalism, profit maximization by the capitalist firm induces competition and that induces technical innovation, as firms look for ways to increase productivity and reduce costs. That in turn increases the productivity of labor, further propelling industrialization. The factory system evolved to capture economies of scale by using the steam engine and organizing large groups of hired labor. Factories replaced artisan shops, home production, and the cottage industry. Thus, it was not just the application of science and technology to production that created modern economic growth but also a new way of organizing production: capitalism. Capitalism uses a market economy based on private ownership by individuals or corporations and private investment. The role of the state is to protect property rights and contracts. The first Industrial Revolution consisted of the mechanization of production

and the use of inanimate power. Joseph Schumpeter (1961) stressed waves of technological innovation financed by the extension of credit as characteristic of capitalism. "Industrial capitalism" was the initial path out of the Malthusian trap in the West.

*Expansion of Agriculture*

W. W. Rostow (1960) wrote that an increase in agricultural output beyond what is required for subsistence is needed to free up agricultural workers for industrial jobs, to supply food, for markets and for capital. Commercialized agriculture tends to produce this increase and thus replaces subsistence farming. One structural change accompanying modern economic growth is the reduction of the percentage of the labor force that is working in agriculture.

*External Causes: Shocks and Substitutions*

Neoclassical economics has had success in explaining the economic growth of countries such as the United Kingdom and the United States. But not all countries were so fortunate as to evolve by market mechanisms. Some countries were awakened by external shocks, from the commercial or military power of those early industrialized states. When a nation is shocked, it can respond to this challenge by adopting the industrial capitalism of its conquerors as Japan did in the nineteenth century, or it can submit and be dominated, as China did. Alexander Gershenkron (1962) wrote that, in backward countries, the state could make substitutions for one of the missing prerequisites for industrial development, such as taking the place of entrepreneurs if none were available in the private sector. He theorized that the more backward the country, the more the state focused efforts on developing heavy industries and large-scale production. This path to modern economic growth is more than usually uneven and dualistic, with the modern sector developing independently of the traditional sector, and often with no increase in the standard of living of the common people, as can be seen in Germany and Russia in the nineteenth century.

*External Causes: International Trade and Investment, Colonialism, Imperialism*

Neoclassical economists argue that the colonial powers broke down the barriers to growth in non-European countries, freeing them up for potential modernization. Furthermore, foreign-promoted export expansion—by means such as developing plantations and mines—enabled non-European countries to grow. Some European colonial powers left additional positive legacies, such as the rule of law, secure property rights, modern transportation, and communication systems. On the other hand, Andre Gunder Frank (1966) argued that in the nineteenth and twentieth centuries, foreign capitalist penetration of countries such as Argentina and India led them to be dependent on the European powers, resulting in underdevelopment at home. This is called dependency theory. On this view, at the time of independence, after draining off the surplus production for decades, the colonial powers left their former colonies in the hands of a local elite uninterested in pursuing economic development, and they left a transportation system designed to facilitate export from the colony to the mother country, rather than for internal development within the country or region. The railroads led from the interior of Africa and Latin America to the ports but not between places within each continent. Technological change brought by the colonial powers did lead to an increase in productivity, but it did not lead to an increase in the standard of living of the larger population, which stagnated during the colonial period.

## Effects of Modern Economic Growth

The increasing productivity of labor achieved under modern economic growth causes the cost of basic goods to fall and enables the real wage to rise. There is also a huge decrease in the drudgery of labor, and life in general becomes more comfortable. A large middle class develops between the old groups of rich and poor of premodern days. Medical advances and improved sanitation lead to declining mortality and morbidity rates (Riley, 2001). All these signify great increases in material well-being of the bulk of the population. North (1981) and Lindert and Williamson (1983) take this view of the consequences of modern economic growth.

Not all economic historians agree on the positive effect of modern economic growth on the working class and on the poor. Frederick Engels (1845/1974) and Eric Hobsbawm (1968) saw this effect as negative. Cynthia Taft Morris and Irma Adelman (1988) argue that the nature of this effect depends on the timing and pace of industrialization, developments in agriculture, and the growth of population. They theorize that, in the early stages of modern economic development, per capita income and the average wage in agriculture and in industry do fall somewhat, and the proportion of the population in extreme poverty does rise. This happens where change was rapid and new employment was not available to replace the traditional jobs that were lost due to economic change. However, Morris and Adelman maintain that in the long run, poverty was reduced by the continued growth in the productivity of labor.

Due to contemporary concerns over climate change, Angus Maddison (2007) has projected trends in the future relationships between economic growth, energy consumption, carbon emissions, and global warming. This theorizing sheds light on pollution as a major effect of the Industrial Revolution. Starting with the use of coal, the first fossil fuel, and later petroleum, the second fossil fuel, the energy sources for industrialization produced the negative externality of pollution. On the other hand, Indur

Goklany (2007) presents arguments and evidence that we are living on a cleaner planet due to technological innovation and economic development.

Modern economic growth gave wealth and power to the West, and that economic power led to political and military power and the ability to dominate much of the world. Some find this course of events to be negative and label it *Western imperialism.* Others interpret it as merely a matter of increased trade and investment of the West in the rest of the world with positive effects for all.

## The Historical Record

### Europe

The economic history of Europe is studied because Europe was the first region of the world to develop modern economic growth and because of its successful offspring: the United States, Canada, Australia, and New Zealand.

To discuss the economic history of Europe as far back as the Roman Empire, we must leave the realm of modern economic growth and enter the realm of long-term economic development. North (1981) stressed the importance to continental Europe of inheriting Roman law, the codification of property rights. The Roman Empire lasted for a thousand years, with a high point in 200 CE. The Roman Empire conquered what today are the European nation-states of Italy, France, Spain, the United Kingdom, and Romania. The Roman Empire collapsed in the fifth century CE. In the Dark Ages that followed, there is little empirical evidence of economic activity in Europe. It is known that the rise of Islam caused the Mediterranean to become a Muslim lake in this period.

Economic historians have analyzed the rise and fall of the European feudal system, with a labor system of serfdom and a political system of decentralized control by lords. Eventually, long-distance trade was revived and cities evolved in previously rural economies. Manufacturing advances such as in the production of woolen yarn and woolen garments took place in northern Italy and Flanders.

Robert Brenner (1987) ignited a debate by arguing that in the fourteenth century, there was a great divide between Western and Eastern Europe. In the West, the feudal system evolved from a labor system of dependent serfs to a labor system of independent peasants, and thus forward economic developments could occur. In the East, the feudal system was confirmed in the form of serfdom and thus doomed to backwardness. This was also the era of the Black Death, in which so many Europeans died that there was a fundamental change in the labor-to-land ratio. In Western Europe, that enabled serfs to become free peasants.

Jan de Vries (1976) wrote a general history of Europe from 1600 to 1750. This is a crucial period because it was a period of crisis in Europe's traditional economy after the great expansion of the sixteenth century and before the great expansion that came with the Industrial Revolution. Countries and regions responded differently as they reached the limits of earlier forms of economic growth. Economic activity and political power shifted from the Mediterranean and the exhausted empire of Spain to the northwestern edge of Europe and its Atlantic coast. Parts of northern Europe achieved highly commercialized agriculture; southern areas struggled to achieve subsistence. Industry moved to the countryside and was restructured as the "putting-out system," also referred to as proto-industrialization. The Dutch added dynamism to trade with ships that could carry great bulk, creating a new high-volume, low-value trade. The British were creating an Atlantic economy with their colonies in North America. Europe was urbanizing with a few cities growing rapidly. In northwestern Europe, governments invested in canals, roads, and coastal shipping. Markets in land, labor, and even some capital markets were developing. All these activities were creating the preconditions for industrialization and huge economic growth to come in northwestern areas of Europe.

Brenner (1987) argued that in the seventeenth century, the United Kingdom pulled ahead as the strongest economy in Europe by developing capitalist farming using commercialized land, capitalist farmers, and wage laborers. This is in contrast to France, which by and large kept a semi-feudal system of peasants and sharecropping—a less progressive form of agriculture.

A large literature has been created on the economic history of the United Kingdom, the Netherlands, France, Germany, and Russia. I present the main outlines of these stories.

#### *The United Kingdom*

In the hundred years from 1760 to 1860, the economy of the United Kingdom was transformed by the Industrial Revolution. This spread to the European continent a generation later. Under the hegemony of the United Kingdom between 1850 and 1914, there was a dramatic spread of modern economic growth throughout much of the world, in individual countries, in some colonies, and in an international economy with global capital markets, railroad construction, and a global cotton textile market. The discipline of economic history has sought to explain why the Industrial Revolution took place in Western Europe and why it began on the small islands of the United Kingdom.

David Landes (1969) documented the leading sectors of the Industrial Revolution in the United Kingdom: cotton textiles and the coal-steam-iron complex. In the cotton sector, technological innovations included the mechanization of cotton spinning and weaving and the use of machine tools to produce textile machinery. In the coal-steam-iron sector, there was the use of coal, the first fossil fuel, which replaced wood (a renewable energy source). New ways of making iron with coke enlarged the capacity to produce iron. The newly invented steam engine replaced watermills, animal and human power, and sailing ships.

Economic historians have advanced many theories of why the United Kingdom was the home of the first Industrial Revolution. First, it had a modern state dating from the seventeenth-century Civil War, which established a constitutional monarchy. Second, the United Kingdom had an agricultural revolution that freed up labor and capital for the industrial sector. Third, it expanded overseas, created an empire, and dominated international trade and finance. The crown jewel of the Empire was India. At first, the United Kingdom imported Indian textiles and tea, but in time, the United Kingdom forced India to import cheap cotton textiles from the United Kingdom, as well as iron. The British used their superior military power to get the Chinese to import opium in exchange for tea. The United Kingdom established an informal empire in Latin America after these nations gained independence from Spain and Portugal. The United Kingdom was a winner in the "Scramble for Africa," thereby gaining a market for its cotton textiles, iron, and railroad building.

There was a highly charged debate in British economic history between the pessimists and optimists on the standard of living of English workers during the Industrial Revolution. The pessimists argue that the impact was negative and that the workers were impoverished. E. P. Thompson (1963) analyzed the making of the English working class, a process that took more than a generation. Engels (1845/1974) described the hard conditions of the working class in England during its second generation. Riley (2001) documented a decline in life expectancy in some industrial cities. Hobsbawm (1968) stressed the terrible insecurity due to cyclical and severe unemployment experienced by three generations of workers. C. T. Morris and Adelman (1988) added the focus on the course of poverty during the Industrial Revolution and demonstrated that in the United Kingdom, poverty increased painfully in early stages, though in time it decreased. The impoverishment of the handloom weavers caused by the introduction of the mechanical loom is an example of how economic development can worsen the condition of parts of the working class, at least temporarily.

The optimists such as T. S. Ashton (1948) argued that the impact was positive and that the standard of living of British workers increased. Lindert and Williamson (1983), mining new sources of accumulating data, demonstrated that from 1820 to 1850, the level of real wages doubled. Gregory Clark's (2007) study led him to conclude that the real wages of urban unskilled workers began to rise by 1815.

Looking back on the debate now, it appears that both sides were right. The first generation of English factory workers was impoverished, but their grandsons reaped the rewards of their sacrifices, achieving a much higher standard of living. Above all, the debate appears to have been about value judgments. The pessimists took the harm to the first and second generations of workers to be important enough to count the effect of industrialization as negative. The optimists took the long-term improvement of living conditions for the majority of workers to count the effect of industrialization as positive, despite the suffering endured by the early generations.

Niall Ferguson (2002) argued that the British Empire was a dynamic force for the good in spreading private enterprise around the world. In addition, he claims that the export of British capital and institutions, such as the common law, a secure land tenure system, and other forms of property rights, was of benefit to the world. The United Kingdom led a world boom of trade and investment from 1899 to 1913, the first era of globalization.

The British Empire had been the largest and most powerful empire in the West since the ancient Roman Empire. However, the world economy eventually found more efficient ways to trade, invest, and grow than by colonizing the non-European world. After World War II, decolonization was accelerated as the peoples of the colonies demanded independence and took action.

### *The Dutch Republic*

In the seventeenth century, before the Industrial Revolution in the United Kingdom, the Dutch Republic was the global power. The Dutch were competitive in international trade because of the sailing vessels they designed that were able to carry large volumes of cargo on transoceanic voyages. They were especially active in the spice trade in the East Indies. In addition, the Dutch Republic was the site of a powerful financial revolution with the founding of a central bank, a national public debt, permanent joint-stock companies, and the Amsterdam Stock Exchange. After 4 years of war between England and the Dutch, the governments of the two great powers divided the East Indies between them, with the Dutch East Indies Company controlling the spice trade in what is today Indonesia and the English East Indies Company controlling the Indian textile trade.

### *France*

The French Revolution got rid of the rent-seeking, landowning feudal class and began the development of democracy in Europe. Yet, the movement of labor from agriculture to industry was much slower in France than in England. French farmers had lower literacy rates and paid a higher percentage of output to direct taxes to the French state. The Industrial Revolution in the United Kingdom was an external shock to the French economy. French industry could not compete with British industry in continental European markets. Challenged by the United Kingdom's development of industrial capitalism, the French took the political route to modernity. In the Second Empire, Louis Napoleon Bonaparte oversaw state-led industrialization, as did the leaders of the Third Republic, established in 1870. The French developed a banking system in Paris to finance the building of railroads in France and Russia. By the late nineteenth century, France was a major industrialized power. The French participated in international trade and

investment in the world economy in the nineteenth century and colonized North and West Africa and Indochina.

*Germany*

Economic history of modern Germany began with Napoleon's defeat of Prussian forces at Jena and the capture of Berlin in 1806. This external shock led to the beginning of the process of unifying the German state. In 1834, a Customs Union, the *Zollverein,* was formed under the leadership of Prussia with most of the other German states. In 1862, Prime Minister Bismarck used the state to industrialize Prussia. The German states were united in 1870 to form Germany. From 1870 to 1914, Germany (along with the United States) created the Second Industrial Revolution based on new technology in heavy industry: chemical, electricity, petroleum, and steel. The Second Industrial Revolution used science-based technological innovation and required the financing and building of fixed capital in the form of plants, machinery, and infrastructure. Large German banks financed the creation of large-scale enterprises with the latest technology. By 1914, Germany was a major industrial power and exporter of capital. Its national output, output per person, and share of world manufacturing were greater than the United Kingdom's. German modern economic growth began later than British and French and was more uneven than is usual in industrial capitalism. The agricultural east of Germany was backward, whereas in the west there had been manufacturing continuously from the Middle Ages, and there was much coal and iron for modern industry. Some economic historians claim that it was the Second Industrial Revolution that created the unparalleled prosperity of the West.

*Russia*

The economic history of modern Russia also began with an external shock, Napoleon's invasion of Russia in 1812. It was not until 1861 that serfdom was abolished in Russia. The freed serfs had to pay for the land they worked, thus keeping them in poverty. Consequently, agriculture was a drag on economic growth, not a handmaiden to it. Russia is a strong case of Gershenkron's (1962) theory that the state can substitute for private entrepreneurs, when they are lacking, and lead an industrialization effort. In 1905, Prime Minister Witte oversaw state-led industrialization. From 1906 to 1914, Prime Minister Stolypin led reform in the agricultural sector, creating private property rights and consolidating small plots into large capitalist farms. More economic progress might have prevented the Russian Revolution in 1917. Central planning with 5-year plans was the strategy of the Soviet communists. In the 1930s, Stalin forced the fastest industrialization of an economy in history, taking 10 years compared to the hundred years it took to industrialize the United Kingdom. This increased the standard of living of industrial workers, but the agrarian workers whose farms were collectivized to feed the industrial workers suffered greatly. The Soviet Union was a case of brutally uneven economic development of late, state-led industrialization.

## Asia, Africa, and Latin America

In Asia, ancient civilizations arose in the river valleys of Mesopotamia, Egypt, India, and China. Human capital development was limited to the spread of literacy in written language among the elites. Physical capital development took the form of irrigation for these agrarian societies. In the medieval era, the rise of Islam from the Arabian Peninsula led to the development of another civilization that spread west along North Africa and east to India and Indonesia. It was spread by Arab warriors and traders. It was a flourishing civilization during the long epoch of the Dark Ages in Europe. The Arabs kept alive the knowledge of the ancient world, which they translated into Arabic. In addition, they translated knowledge from China and India into Arabic. The Islamic Empire covered an area greater than the Roman Empire. Later, Turkish Ottoman invaders from central Asia shocked and then reorganized much of the Middle East. The Ottomans took control from the Arabs but adopted their religion of Islam.

Civilizations also developed beyond the Eurasian land mass: in the New World, the Aztecs and Mayans in Mexico and the Incas in Peru. In sub-Saharan Africa, there arose empires on the plains of Ghana, kingdoms of central Africa, and the states of southern Africa, such as the Great Zimbabwe.

During the "Age of Discovery," from the early fifteenth century to the early seventeenth century, Europeans explored the non-European world, crossing the seas in search of gold, silver, and spices. The Europeans were blocked in the East by Islamic empires and thus could not use the ancient overland "silk road," or the sea route through the Red Sea or the Persian Gulf, to obtain the luxury goods of the East. Facing this challenge, the Portuguese invented the carrack and the caravel, ships that could sail on the open Atlantic. Portuguese explorers rounded the cape of Africa and sailed to India. The Spanish sent Columbus to find Asia by sailing west. In time, the Spanish conquered the ancient empires of the Aztecs in Mexico and the Incas in Peru, stealing their gold and silver. In 1493, the pope divided the world in half along meridians of longitude in the Atlantic and the Pacific, giving Portugal Brazil and all non-European lands to the east of the Atlantic meridian and Spain all the land to the west, including Central and South America and the Philippines. The Spanish sent Portuguese explorer Magellan to sail west to find the Spice Islands (today Indonesia). His ship was the first to circumnavigate the earth. His sailors found the Strait of Malacca, which connects the China Sea with the Indian Ocean. The Portuguese were the first Westerners to reach and trade with Japan. In time, the Spanish and

Portuguese were overtaken by the Dutch, French, and English who ignored the pope's division of the world. Here, then, world economic history begins.

The "Age of Imperialism" saw great rivalry among the industrial nations of Europe, much of which played out in conquering or dominating non-European lands with guns, trade, and investment. The Ottoman Empire (1299–1922 CE) controlled the Middle East, blocking European penetration there until World War I.

Economic historians have tried to explain why Europe pulled ahead of the rest of the world in the modern era. Mokyr (2002) argued that the Industrial Revolution occurred in Europe because of the scientific revolution in the seventeenth century and the Enlightenment in the eighteenth century. He contends that it was the resulting superior technological creativity and knowledge of the British, Germans, and French that enabled them to industrialize first (Mokyr, 1990). Eric Jones (2003) argues that Europe in the early modern era (1400–1800) was developing technology and markets, discovering new lands, and developing a system of nation-states that propelled it to worldwide power and prosperity. He writes that China was a huge empire and controlled from above, and thus it lacked the system of nation-state competition which propelled the great breakthroughs of the Industrial Revolution in Europe. Another disadvantage that China had, according to Jones, was that in the fifteenth century, it closed itself off from maritime exploration and trade, great engines of growth for Europe. Jones also contends that Europe had the advantage of being far from the nomadic raiders from Central Asia that interrupted the trajectory of development in the Islamic Middle East (the Ottomans), India (the Mughals), and China (the Manchus). Recently, Kenneth Pomeranz (2000) has argued that the two causes of Europe escaping the Malthusian trap, led by the United Kingdom, were the colonization of land in the New World and the lucky geographical accident of having coal.

Centuries of commercial capitalism in Europe prior to the Industrial Revolution were part of the advance of Europe over Asia. Thus, the dating of the diverging paths of Europe and Asia becomes an issue. Pomeranz (2000) has claimed and presented some evidence that Europe did not pull ahead of China economically until 1800. This is highly debated. Many economic historians stand by Landes (1969), who demonstrated that from 1500 to 1700, Europe was pulling ahead of Asia in economic growth, and by Maddison (2007), the great constructor of premodern economic statistics of the world, whose statistics showed Western Europe growing twice as fast as the rest of the world from 1000 to 1500 and that Western hegemony was established between 1820 and 1870.

Brief sketches of five major non-European countries and regions are presented below, plus what economic historians had achieved so far in the explaining the economic growth in the non-European world.

### *Japan*

Japan was the first non-Western nation to achieve modern economic growth. Like the British Isles, the Japanese archipelago lies off the Eurasian land mass. This was a time of the crucial importance of water transportation. The external shock that woke up Japan was the gunboat diplomacy of Admiral Perry, who with four U.S. warships demanded that Japan open up to trade. Prior to this shock, Tokugawa Japan (1603–1868) was an isolated, preindustrial feudal society. However, it was not stagnant. The city of Edo (modern-day Tokyo), the seat of the Shogun (secular rulers), in the eighteenth century, was perhaps the largest city in the world. Financing and marketing rice production and coastal shipping developed slowly. To cope with the Western challenge from Perry, the rulers of Japan led a campaign to industrialize and modernize Japan. Feudalism was abolished and a Western-style legal system put into place. These were called the Meiji Reforms. The Meiji Restoration (1860s) gave Japan an emperor but a constitutional monarchy. With a favorable international economic environment, modern businesses arose in the 1880s, and there was a take-off in the 1890s to sustained modern economic growth. Japanese leadership was open to and borrowed methods and ideas from abroad. C. T. Morris and Adelman (1988) describe the growth of productivity in Japanese agriculture from the sixteenth century, leading to market-oriented farmers who emerged from the disintegration of medieval farming in the nineteenth century. They contend that the slow commercialization during the Tokugawa period produced less extreme poverty in Japan than in Russia, India, or China. In the later nineteenth century, the Japanese government had a policy of locating industry in rural areas to absorb underemployed labor, thus reducing poverty. After World War II, Japan rose to become the second largest economy in the world, after the United States.

### *China*

China had a large population on the east end of the Eurasian landmass, just as Europe had on the west end. China manufactured much, invented new technologies, had high levels of literacy, administered exams for civil servants, and had a national market. China had a single government controlling a huge territory for two thousand years. The Chinese invented many important tools but did not apply them economically, such as gun powder. Mokyr (1990) argues that in 1300, China was the site of dramatic technological creativity and yet lost that creativity after this.

In the nineteenth century, European traders forced the Quin Dynasty (Manchu China, 1644–1911) to open to trade, forcing them to take European imports, especially opium. European merchants forced their way into ports such as Shanghai, which became an international city. The British took control of Hong Kong. Chinese economic growth

stagnated in the nineteenth century. The Chinese did not react to the European challenge as the Japanese did; they did not adopt European technology. They exported tea and silk and imported British cotton goods, providing a market second in size only to British India.

The Republic of China, formed in 1912, ended two thousand years of imperial rule and started the early stages of industrialization. This was interrupted by the Japanese invasion in the 1930s and the Chinese Communist takeover in 1949. Communist central planning with 5-year plans led to disasters in agriculture and to forced attempts at industrialization. However, through public investment in health and education, there was an impressive increase in the life expectancy of the Chinese people, a rise in literacy rates, and a decline of the proportion of the population in absolute poverty.

Since 1978, China has been taking off into sustained modern economic growth by opening up to global economic forces and allowing economic competition internally. The world has seen unprecedented economic growth rates in China, raising standards of living and lifting hundreds of millions of people out of poverty.

R. Bin Wong (1997) challenged Western economic historians by documenting "Smithian" economic growth (from trade, not technology) in China up to the nineteenth century. He argues that the West did not overtake China until the Industrial Revolution and its use of inanimate energy. The causes were more political than economic—while China was a huge empire, Europe was a system of competing nation-states propelling modern economic growth.

### *India*

K. N. Chaudhuri (1990) presented the Indian subcontinent as lying at the center of a huge sea trade system with Western Asia, across the Arabian Sea and along the Persian Gulf and the Red Sea, and with Eastern Asia via the Bay of Bengal and through the Straits of Malacca to the China Sea. Indian and other Asian merchants such as the Arabs were the leading economic agents in Asian development for a thousand years before the arrival of the Europeans.

Jones (2003) argues that the economic development of India was interrupted by the invasion of the Mughals from Central Asia in the sixteenth century. In the nineteenth century, India faced a second shock with the invasion of European merchants and in time European governments. British India (1857–1947) has been analyzed by British economic historians. Morris D. Morris (1960) argued that economic growth in India in the nineteenth century was constrained by lack of productive capacity, not by the politics of colonialism. Under the 90 years of British colonialism, India was open to trade and investment from the world economy, connected to it by the Suez Canal, railroads, and telegraph. The British developed land markets and established secure property rights. They commercialized agriculture and developed export crops of cotton and tea. Much of this was accomplished by coercion and violence. Tirthankar Roy (2002) argued that, in the first 60 years of British colonialism, there was economic growth in India and an increase in the standard of living. But after World War I, conditions worsened for the whole world economy and thus for India, which was connected to that world economy.

For the postcolonial period, Indian scholars tended to focus on the negative consequences of a century of British rule in India. The British certainly did leave India with low life expectancy, low literacy, stalled industrialization, and trouble feeding itself. It is not known whether this was worse than they found it in terms of the masses of the Indian population. From 1947 to 1990, the leaders of independent and democratic India closed the economy off from the West. Progress was made domestically using central planning with 5-year plans focused on agriculture. With the help of Western science and philanthropy, the Green Revolution of using high-yielding hybrids of rice and wheat after 1965 enabled India to feed its growing population. Indian central planners practiced import-substitution (protecting infant industries and producing at home what was previously imported) and used subsidies to support industrialization as well as maintain the cottage industry. Their goal was equity as much as efficiency.

Since the 1990s, the Indian government has been liberalizing the economy and opening it to the world. Currently, India is experiencing modern economic growth due to the expansion of the information technology sector and other business services and, to a lesser extent, labor-intensive manufacturing. Economists are studying the transformation from the lower "Hindu" rate of growth under central planning to the accelerated rate under economic reforms. India still lags behind China in educating its people and in investment in infrastructure.

### *Latin America*

While there are many histories of the Aztec, Mayan, and Incan civilizations, there have been few studies of their economic histories. In English, there are some economic histories of colonial Latin America. Most of the countries in Latin America gained their independence from European powers by the early nineteenth century: Brazil from Portugal and Argentina and the rest of Latin America from Spain. The empirical research of C. T. Morris and Adelman (1988) on the role of foreign economic dependence in the nineteenth century found that while Argentina was politically independent, an alliance of indigenous landlords and the British investors favored exports (beef and wheat) over domestic development. Argentina became heavily dependent: Domestic growth was dominated by their exports at the expense of developing a domestic market, foreigners dominated trade and banking, and investment was financed by foreigners, mainly the British. Morris and Adelman found Brazil to be moderately dependent on foreigners, first for the exportation of sugar and later coffee. Also, slavery was practiced in Brazil until

1888, which was not favorable to modern economic growth. Slavery generates no incentives for technological innovation because the cheap labor of slaves is available. Slavery did produce wealth for the slave owners, but it did not create households that had effective demand for consumer goods because slaves were kept at a low standard of living. Perhaps even more important for economic development, slave owners did not have demand for producer goods because they could not trust their slaves with more than rudimentary tools.

Enrique Cardenas and his colleagues published three volumes on the economic history of Latin America (Cardenas, Ocampo, & Thorp, 2002), in which they challenged the views of dependency theorists (Cardoso & Faletto, 1979; Frank, 1966) on the negative effects of European colonization and imperialism. They present evidence that, from 1870 to 1930, in addition to the expansion of foreign-promoted exports from Latin America, there was also development of domestic markets and manufacturing capacity. They cite the technological advance of the steel-hulled steamship, which reduced international transportation costs and thereby enabled Latin American countries to export their mineral and agricultural raw materials. It is true that foreigners invested in these sectors, as well as financed and built the railroad system needed for export, but in this view, that dependence was not negative but helpful. There is evidence of a domestic market made up of wage workers from the ranches, plantations, mines, railroads, and ports who wanted locally produced cotton clothing, beer, and cigarettes. During the Great Depression, industries in Latin America were able to grow domestically as world trade collapsed. Unlike the Washington consensus view that export-led growth is the best strategy for development, Cardenas et al. (2002) argue that import-substitution industrialization made sense in the 1940s to the 1960s for the development of national economies. In this strategy, a nation imported capital goods and then produced consumer goods for the domestic market that otherwise would have had to have been imported.

### *Africa*

There were many complex African empires and civilizations before the Europeans came to Africa. There are some histories of European colonization of Africa. But the development of the economic history of Africa as a discipline is just beginning.

From the sixteenth to the nineteenth centuries, about 9 to 12 million enslaved Africans were brought to the New World, mostly to Brazil. Seventy percent of all slaves were used on sugar plantations. European competition drove the process, starting with the Portuguese, the Spanish, and, later, the slave and sugar merchants of France, England, and Holland. By the eighteenth century, the British were the leading slave traders. The triangular trade consisted of Europeans taking copper, cloth, guns, ammunition, and alcoholic beverages to West Africa, exchanging them for African slaves, who were taken via the Middle Passage to the West Indies, where the slaves were traded for sugar, rum, molasses, and tobacco, which were then sent to European markets.

Eric Williams (1966) set off a debate that has continued for over a half century on the role of profits on the slave trade, as well as on slave production in the New World, for the economic development of Europe. He argued that Europe could not have taken off into modern economic growth without those ill-gained profits. Pomeranz (2000) presented empirical evidence showing that the amount exploited from this brutal activity was not as large as the profits made in the United Kingdom and Europe from their own domestic production of farms, workshops, and factories. The implication is that Europe could have advanced without the profits of the slave trade and slave production. In 1807, the United Kingdom abolished the slave trade, and in the 1880s, Brazil abolished slavery, the last country in the Western Hemisphere to do so. Thomas Pakenham (1991) has documented the "Scramble for Africa," when, from 1876 to 1912, the European nations divided sub-Saharan Africa among themselves for needed raw materials and for markets for their manufactured goods.

## Policy Implications

Economic history of the developed world can be used by development economists searching for the path to modern economic growth for those nations not yet on their way. Development economists are interested in poverty alleviation in addition to the goal of modern economic growth. One lesson for poverty reduction seems to be that agriculture and the rural sector need to be developed. Another implication is that the role of the state is complex in promoting modern economic growth. There is a role—to provide public infrastructure and human capital—but there is the danger of too much government intervention blocking market forces for change. National leadership and political will are needed to lead the drive for modern economic growth. The historical record is mixed on whether a nation should use an export-led growth strategy or an import-substitution industrialization. The case of Japan shows how policy can prevent an aggravation of poverty during industrialization. The cases of China and India suggest that nations need to be open to the world economy.

## Future Directions

Each generation has to write its own history, its own interpretation of past events, because it is faced with new problems that need different lessons from history. The twentieth century is now history. One possible agenda for finding out how to promote modern economic growth in the poor nations of the world would be to compare the Industrial Revolutions of the United Kingdom and continental Europe with that of the Soviet Union in the 1930s and with

China in the last quarter of the twentieth century. The continuing power of European and American banks and industrial firms in Africa needs to be analyzed and compared with the internal constraints on modern economic growth of the intrusive governments of many of the nations of Africa. Globalization now under the United States should be compared to globalization in the nineteenth century under the United Kingdom. Just as economic historians have tried to explain the rise of the West, now the rise of the East should be analyzed. In addition, economic historians should benefit from researching the economic history of the Middle East.

## Conclusion

The neoclassical model of growth explains the rise of the United Kingdom, France, and the United States in the nineteenth century, but heterodox models are needed to explain the underdevelopment of nations such as Argentina and India in the nineteenth and twentieth centuries. Dependency theory is still relevant but must be balanced with the neoclassical economic historians' view that imperialism (international trade and investment) was the pioneer of capitalism in the non-European world. The benefits of capitalism and imperialism are an increase in the material standard of living and an increase in life expectancy for many people. The costs seem to be a widening divergence in the fate of rich and poor countries and the pollution of the planet.

## References and Further Readings


Ashton, T. S. (1948). *The Industrial Revolution, 1760–1830.* Oxford, UK: Oxford University Press.

Becker, G. S. (1975). *Human capital.* New York: Columbia University Press.

Brenner, R. (1987). Agrarian class structure and economic development in pre-industrial Europe., In T. H. Ashton & C. H. E. Philpin (Eds.), *The Brenner debate* (pp. 10–63). Cambridge, UK: University of Cambridge Press.

Cardenas, E., Ocampo, J. A., & Thorp, R. (2002). *An economic history of twentieth century Latin America.* New York: Palgrave.

Cardoso, F. H., & Faletto, E. (1979). *Dependency and development in Latin America* (M. M. Urqudi, Trans.). Berkeley: University of California Press.

Chaudhuri, K. N. (1990). *Trade and civilization in the Indian Ocean: An economic history from the rise of Islam to 1750.* Cambridge, UK: Cambridge University Press.

Clark, G. (2007). *A farewell to arms: A brief economic history of the world.* Princeton, NJ: Princeton University Press.

de Vries, J. (1976). *The economy of Europe in an age of crisis.* Cambridge, UK: University of Cambridge Press.

Engels, F. (1974). *The condition of the working class in England in 1844* (F. K. Wischenwetzky, Trans.). St. Albans, UK: Panther. (Original work published 1845)

Ferguson, N. (2002). *Empire: The rise and demise of the British world order and the lessons for global power.* New York: Basic Books.

Frank, A. G. (1966). *The development of underdevelopment.* New York: Monthly Review Press.

Gershenkron, A. (1962). *Economic backwardness in historical perspective.* Cambridge, MA: Belknap.

Goklany, I. M. (2007). *The improving state of the world.* Washington, DC: Cato Institute.

Hobsbawm, E. J. (1968). *Industry and empire: An economic history of Britain since 1750.* London: Weidenfeld & Nicolson.

Jones, E. (2003). *The European miracle: Environments, economies and geopolitics in the history of Europe and Asia.* Cambridge, UK: Cambridge University Press.

Kuznets, S. (1968). *Toward a theory of economic growth.* New York: W.W. Norton.

Landes, D. S. (1969). *The unbound Prometheus: Technological change and industrial development in Western Europe from 1750 to the present.* Cambridge, UK: Cambridge University Press.

Lindert, P. H., & Williamson, J. G. (1983). English workers' living standards during the Industrial Revolution: A new look. *The Economic History Review, New Series, 36,* 1–25.

Maddison, A. (2007). *Contours of the world economy 1–2030 A.D.: Essays in macro-economic history.* Oxford, UK: Oxford University Press.

Malthus, T. (1993). *An essay on the principle of population.* Oxford, UK: Oxford University Press. (Original work published 1798)

Mokyr, J. (1990). *The lever of riches: Technological creativity and economic progress.* Oxford, UK: Oxford University Press.

Mokyr, J. (2002). *The gifts of Athena: Historical origins of the knowledge economy.* Princeton, NJ: Princeton University Press.

Morris, C. T., & Adelman, I. (1988). *Comparative patterns of economic development, 1850–1914.* Baltimore: John Hopkins University Press.

Morris, M. D. (1960). Report on the Conference on Asian Economic History. *Journal of Economic History, 20,* 435–440.

North, D. C. (1981). *Structure and change in economic history.* New York: W. W. Norton.

Pakenham, T. (1991). *The scramble for Africa.* New York: Avon Books.

Pomeranz, K. (2000). *The great divergence: China, Europe, and the making of the modern economic world.* Princeton, NJ: Princeton University Press.

Ricardo, D. (1819). *The principles of political economy.* London: John Murray.

Riley, J. C. (2001). *Rising life expectancy: A global history.* New York: Cambridge University Press.

Rostow, W. W. (1960). *The stages of economic growth.* Cambridge, UK: Cambridge University Press.

Roy, T. (2002). Economic history and modern India: Redefining the link. *Journal of Economic Perspectives, 16*(3), 109–130.

Schumpeter, J. A. (1961). *The theory of economic development* (R. Opie, Trans.). Oxford, UK: Oxford University Press.

Smith, A. (1976). *An inquiry into the nature and causes of the wealth of nations.* Chicago: University of Chicago Press. (Original work published 1776)

Solow, R. (1970). *Growth theory.* New York: Oxford University Press.

Thompson, E. P. (1963). *The making of the English working class.* London: Penguin.

Williams, E. E. (1966). *Capitalism and slavery.* New York: Capricorn Books.

Wong, R. B. (1997). *China transformed: Historical change and the limits of European experience.* Ithaca, NY: Cornell University Press.

# 3

# Economic Methodology

Timothy A. Wunder
*University of Texas at Arlington*

Undergraduate students are often introduced to economic concepts using graphical expressions, and the expressions become the defining method of explaining and exploring the economic realm. The graphs become what the students perceive as "economics," and this is what is meant as "methodology." If you ask the students, "How do you do economics?" the answer would be based on the graphical examples offered in the classes. Methodology is a complicated way of saying, "These are the tools economists use to explain the economy." Beginning students are offered very simple tools to explain the concepts of supply, demand, and equilibrium, but the ideas, as well as the graphs, are all part of the methodology.

In most undergraduate economics classes, theory is offered to the student as a body of cohesive ideas set within a structure that seems internally consistent and mutually reinforcing. Such a monolithic view of theory may offer simplicity to a student who is just learning the basics, but this monolithic vision of methodology offers a false level of certainty based on an oversimplified version of ideas. Methodology used by economists today is very different from what was used 30 years ago, and what was used 30 years ago was very different from the methodology used by such great economists as Adam Smith and David Ricardo. The economic methodology that is taught in undergraduate courses today is the result of centuries of intellectual debate, and the origin of this body has been filled with differing thinkers often in violent disagreement with each other. It is in this history of methodology that the origins of undergraduate theory can be found, and an exploration of this history is both exciting and illuminating.

There are generally two standard bodies of theory that are offered in undergraduate economics classes. The microbody of theory offers a vision based on the individual behavior of consumers and firms. These actors operate under a theory of rational self-interest that leads to stable social outcomes referred to as equilibriums. The second body of theory is the macroexploration of economic activity. In this body, the economy as a whole is examined, and differing reasons are offered to explain why certain events occur the way they do. Undergraduate classes suggest that the microeconomy is ruled by the prevailing forces of supply and demand, and the macroeconomy is explained by aggregate supply and aggregate demand. This chapter's focus will be to explore where these theories and this methodology arose from.

## Present State of Economic Methodology

Trying to describe what is the present state of methodology in economics is a lot like trying to summarize modern culture. Whatever statements are made are going to be overgeneralizations with respect to differing groups. The methodology of undergraduate economics and that of professional economists is very different. The undergraduate will often learn about economic theory using graphs and some math, and even a little econometrics may be thrown in. This type of methodology was the prevailing form done by professional economists perhaps 30 years ago, but it is vastly different from what is done by professional economists today.

Professional economics is in a state of transition with respect to the methodology being used. The methodology

taught in undergraduate classes, however, does offer a reasonable, though simplified, vision of methodology that was used by most professional economists throughout the majority of the twentieth century. In the past, there were four major components to professional economic methodology. First was that economists agreed that the study of economics was based on the study of the individual, not of groups. Basing the study on the individual is known as methodological individualism, as opposed to methodological collectivism, in which individuals are studied within the context of the groups to which they belong. Second, economists held a mechanistic vision of the economy where economic laws were like physical laws seen in Newtonian physics. Equilibrium was the ultimate outcome that the economy would eventually revert back to. Third, mathematical rigor and deductive reasoning were used in place of empirical observations because there is an inability to conduct controlled economic experiments. Finally, internal consistency of theory was more important than empirical evidence, and evidence contradicting consistency was not highly valued.

As this chapter is being written, professional economists are questioning the appropriateness of all four methodologies. First, many within economics are now calling into question the appropriateness of studying economics based on solely individual action. Some economists, such as behavioral economists, are exploring how individuals behave within group economic settings, and the research is becoming popular. Readers interested in an introduction to this type of economics should read *Predictably Irrational* by Dan Ariely (2008). Second, the mechanistic concept of an economy headed toward an ultimate equilibrium is also being called into question, and this can be seen in books such as *The Origin of Wealth* by Eric D. Beinhocker (2006). Third, economists are not abandoning math, but deductive reasoning is giving way to more inductive methods. Under deductive reasoning, economists would state certain assumptions they believed to be true; by deductive logic, if the assumptions were true, then the conclusions would have to be true. Modern methodology is becoming much more inductive in that economists are testing theories using econometric techniques. This inductive method essentially starts by observing what is going on in the world and then hypothesizing why it is going on. The biggest unwritten rule in modern professional economics is that theory should be explored using data sets and econometric techniques, and this is far more inductive than methodology from even 30 years ago. Finally, the idea that internal consistency is more important than empirical verification is clearly falling by the wayside. The goal of many new economic studies is to use econometric techniques and data to either prove or disprove certain aspects of theory. For information on this issue, see Colander (2000, 2005, 2009); Colander, Holt, and Rosser (2007–2008); and Davis (2007). Modern methodology is in a state of flux, so this chapter will explore the origins of the methodology that was predominant 30 years ago and is still taught as "economics" in undergraduate classes today.

It should be noted that not everybody agrees with the above interpretation of the current state of methodology. Some historians of economic thought would disagree with this description of modern methodology and would disagree about the openness of economics to differing theories. In fact, many economists call themselves heterodox economists who would argue that the changes listed above are superficial at best. Readers interested in examining some materials from these heterodox thinkers can find many online resources at www.heterodoxnews.com. The truth is that it is difficult to determine who is correct because gradual change is much harder to recognize than is radical change. Often, paradigm changes in a discipline are radical in nature, and two periods in time are readily identifiable. Thomas Kuhn (1970) advanced this idea in *The Structure of Scientific Revolutions,* and the idea has been readily adopted by many who study the history of economic thought. The changes occurring in the methodology of economics today seem transformative of the discipline, but they have not come as an abrupt disjuncture from the previous methods. With that note, the chapter now turns to the origins of modern economics starting more than 200 years ago.

## The Classical Paradigm

In 1776, Adam Smith wrote *An Inquiry Into the Nature and Causes of the Wealth of Nations,* and economics as a modern intellectual discipline began. Most of what Adam Smith wrote in *Wealth of Nations* had been written about previously, but Smith brought all of the disparate ideas together into one work. With the unification of these ideas into a single work, a new and unique way of analyzing economic activity had been created. Yet what Adam Smith wrote in 1776 and what students are taught in undergraduate classes today are very different. In fact, Smith created a new methodology for exploring economic topics, but that methodology is not the methodology used today. This section will first explain the accomplishments of Smith and then talk about the methodology known as classical economics.

### Smith's Work

Smith's major contribution was to summarize a large body of thought into a single work, but Smith also explained how an economic system could work when there was no central control to direct the system. Smith argued that an economic system could operate without any form of central guidance, and society could be very well off allowing such a system to operate. Smith recognized that wealth was based on the ownership, use, and construction of goods and services rather than the possession of money. Smith showed how an economic system

based on people acting in their own rational best interests could lead to a socially desirable outcome. Anyone taking a modern introductory economics class can recognize these ideas in modern theory.

Adam Smith wrote *Wealth of Nations* to refute some of the mercantilist ideas that were prevalent at the time. Mercantilist policies emphasized the accumulation of gold to sustain military strength. Smith recognized that the end goal was the production of the materials needed to support the military. It was not the gold that supported the military; rather, it was the ability to produce food, ships, guns, and other items that ultimately allowed for maintained military strength. Modern economists are still far more interested in the exploration of the production of real goods and services than they are interested in the issues of money. The exploration of money in the realm of modern economics is almost always associated with how money affects the production of goods and services.

This materialist vision of economics was a breakthrough, but the idea that an economy could operate without centralized control was an even bigger revolution. As far back as the ancient Greeks, there was a negative connotation to people operating in an economic sphere with the sole purpose of gain. Aristotle and Aquinas, as well as most other previous thinkers, suggested that trade for profit was somehow unnatural. Smith confronted this paradigm, and his arguments shifted social thought in a completely new direction. Smith argued that individuals took the actions they did because they were pursuing their own self-interest. The baker makes loaves of bread for others to use in sandwiches but not out of generosity. She bakes so that she can sell the bread and use that money to buy the objects she wants. It is true that the baker serves the sandwich eater, but she does not do so out of charity; rather, she does so in pursuing her own self-interest.

Smith (1776/1965) argued that self-interest would direct most people's activities toward ends that would benefit society as a whole. He argued that an individual working in pursuit of his own self-interest "intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention" (p. 423). The pursuit of self-interest alone could lead to outcomes that would benefit society in general. This was unprecedented in previous economic thinking, and this idea would become a cornerstone of economic methodology from this point forward.

## Ricardo and Malthus

Smith's (1776/1965) work became a watershed moment in economic thinking. By offering a system of thought that demonstrated how a decentralized economy worked, he allowed others to explore some of the components of such an economy. There were many important thinkers involved in the classical period, but the two who take on the greatest significance with respect to shaping future methodology are Thomas Malthus and David Ricardo. It is interesting to note that these two thinkers were in dramatic disagreement about many of the important classical ideas.

Thomas Malthus was an Anglican curate who began his exploration of economic topics to demonstrate the futility of utopian experiments to improve the lives of the poor. Malthus is most famous for his population doctrine that stated food production increased at an arithmetic rate, whereas human population increased geometrically. Given these differing rates, human populations would be forced to subsistence living. Malthus believed that as food grew scarce, relative to population, only two options were available: Either human population had to control birth rates, or nature would increase death rates. Malthus did not discuss issues such as birth control, so the only method of population control was human abstinence, and he had little hope that human self-control would win out. The thinking behind this population doctrine still is influencing modern discussions and also influenced thinkers outside economics, including Charles Darwin.

Related to this population theory was a concept that became known as the wage fund doctrine. In this theory, Malthus proposed that the ability of capitalists to fund projects was based on the available food in the system controlled by the capitalist. In the end, the workers had to be paid enough so that they could feed themselves and perpetuate the next generation. Malthus argued that the behavior of workers was such that increasing the wages of the workers would lead to increased procreation. Increased procreation would lower the pay of workers in future generations. In the end, the only stable outcome was one in which the wages paid the workers were at a subsistence level. The overall outcome in any economy might show temporary improvements in the material condition of the population, but in the long run, the condition would revert to a steady-state subsistence level.

David Ricardo was born into a merchant family and was able to earn substantial wealth as a stock broker before writing on economics. In 1817, and republished in several editions, Ricardo published *The Principles of Political Economy and Taxation,* which replaced Smith's *Wealth of Nations* as the primary reference for economic thinking. In *Principles,* Ricardo moved away from the Smithian methodology of deductive reasoning with inductive analogies to demonstrate the ideas. Ricardo moved into a very formal method of economics that was strictly deductive in manner. In this method, referred to as deductive reductionism, abstractions became a cornerstone of theory, and simplifications allowed for the creation of a model from which generalizations could be drawn. Ricardo very rarely started out with observations in the real world; instead, he began with his theories and proscribed advice for real-world situations based on how the model predicted the real world would react.

Ricardo created these models because he was interested in exploring why income distributions changed in the manner they did. To explain how incomes were determined required an explanation of why certain items traded

for higher prices than others. Inevitably, any such exploration will eventually call into question what is the source of value in a system, and answering that question is much harder than it may seem. Ricardo offered up a theory of value that took into account rents going to landowners, profits going to capital owners, and wages going to workers. However, Ricardo's theory of value really emphasized the absolute importance of labor in the whole process. Most historians of thought agree that Ricardo's theory of value was very much based on labor's role in the productive process, and thus it became known as a labor theory of value. An example to explain this idea is simple. In a primitive society where it takes 4 hours to capture a deer and 2 hours to capture a beaver, two beavers will trade for one deer.

Ricardo and Malthus were contemporaries and were in correspondence with each other for an extended period of time. They influenced one another's work, and in some cases, they were also very much in opposition to each other's ideas and methods. One of the greatest conflicts between these two thinkers involved the work of French economist J. B. Say. Say argued that a decentralized economy, as described by Smith, would by its very nature always use the whole of the resources available. This idea has come to be known as Say's law and has been abbreviated to read "supply creates its own demand." Suppose a baker sells a cake for $30; the baker immediately has created $30 worth of purchasing power. The process of creating a product immediately creates the purchasing power in the system to buy goods of equal value. If the baker does not spend the purchasing power herself, then she will lend the money to someone who will. On an economy-wide basis, this suggests that although there might be an overabundance in one good, there cannot be a general overabundance (known as a recession).

Malthus and Ricardo found themselves on differing sides of this issue. Ricardo was a backer of Say's law, whereas Malthus argued that the law was flawed. Ricardo believed that the Say's law was deductively true and therefore should be believed, whereas Malthus simply pointed to the history of continuing recessions. Ricardo believed that any general underconsumption (i.e., recession) could only be temporary and therefore dismissed it as a topic. Malthus believed underconsumption could occur for an extended time period and thought that looking at the issues involved was important. In the end, Ricardo's views became dominant until the arrival of Keynes, who will be discussed below.

## J. S. Mill and the Decline of the Classical Paradigm

The classical paradigm started in 1776 with *Wealth of Nations* and came to an end somewhere in the late 1800s. Although a precise final date cannot be given, a final great thinker of the movement can. John Stuart Mill was the son of James Mill, who was also a classical economist. J. S. Mill had a rigorous childhood education, and it is probable that he was one of the finest minds of his time. He was able to bring together all of the pieces of the classical paradigm and polish them in a way that established the high watermark of that school of thought.

Mill both explained the classical school's thoughts and had within his works the seeds of the economic thinking that would come to replace the classical paradigm. Mill was a firm believer in Say's law and a follower of the Ricardian tradition of deductive reductionism as a substitute for the inability to perform experiments in economics. Mill was a proponent of the separation of positive economics from normative economics, and he argued that economists could separate explanations of how the economy works (positive economics) from the moral validity of the economic outcomes (normative economics). This positive–normative distinction is still highly influential in economic circles today. Most economic theory tries to explain the economy from a positive perspective, and most economists will try to make their normative opinions known. However, other economists, such as Joseph Schumpeter, have suggested that the normative and positive aspects of economic theory are not so easily separable. Mill believed by concentrating on positive economics, the discipline could become more scientific in method and economic understanding could be improved.

Probably the most important aspect of Mill's contribution was his ability to summarize and defend the classical position against the onslaught of opposing social thought at the time. The mid-1800s was a time of great social upheaval, and there was a large reactionary movement against the social changes imposed by the rise of industrialization. Because the classical paradigm was seen by many as a defense of the capitalist system, Mill's ability to maintain the school's prominence was noteworthy. He was able to do this partially because he sympathized with the beliefs of the system's critics. Mill was able to refine the classical theory of economic growth and argued that the stationary state, which was a major prediction of classical theory, was not necessarily a bad outcome. Mill believed that the arrival of the new stationary level of economic activity could be achieved within a context of a more just social distribution of income. Mill thought differently than Malthus, who argued the future stationary system would condemn the multitudes to subsistence. Mill disagreed with Ricardo, who argued that the wage fund was an unchangeable fact that demonstrated that any attempt to raise workers' wages was doomed to failure. Instead, Mill argued that a level of social justice was possible within any future stationary system.

In defending the classical system, Mill laid the foundations of many ideas that were to come to dominate postclassical economics. Within Mill's work can be found the seeds of a supply-and-demand explanation of value as a replacement for the labor theory of value. Mill had nascent discussions on general equilibrium economics in his writings. His separation of positive from normative explanations came to

play a major role in future economic thinking, as did his separation of production from distribution. In the end, Mill was the greatest classical thinker, but he also was the economist who began the end of the classical paradigm when he refuted the belief in the wage fund doctrine. In Mill's writings, the highest point of classical theory can be found. In these same works are the seeds of thought that would come to replace the classical school with a new paradigm known as the neoclassical school.

### Objections to Classical Thought

The history of economic methodology is a history of competing ideas. Although the classical school has been offered as the economic methodology of the period, it should be noted that there were competing thinkers who disagreed with the methodology of the classical school. It is important to mention some of the competing thinkers before moving on to the neoclassical paradigm.

### Non-British Thought

The classical theory of economics was heavily influenced by the work of British thinkers, and it should be noted that British philosophical tradition is very different from other traditions. The British tradition emphasizes the importance of individualism as the foundation of social organization, whereas thinkers on the European mainland had a far more social bent in their ideological perspective. In part, it is this difference in the importance of individualism that is a defining difference in methodology between the classicals and their detractors.

Some of the first intellectual criticisms of the classical perspective came from a group of thinkers who have been labeled utopian socialists. Utopian socialist thought differed greatly from the socialist thought of Karl Marx, who will be discussed below. The utopian socialists included individuals such as Claude Henri de Saint-Simon, Robert Owens, Charles Fourier, and many others. The works of these thinkers all emphasized the need for a more social view with respect to economic methodology. The classical perspective emphasized the importance of individual self-interest as a central component to an optimal organization within an economic system. The utopian socialists argued that the capitalist system created great disharmony between differing classes, and neglecting to recognize social disharmony was a major flaw in classical methodology. In their works, the utopians emphasized the importance of planned activities in the economic sphere.

A second group severely critical of the classical methodology was the historical school. The historical school argued that one of the primary premises of classical methodology, separating theory from social context, was flawed. Remember that it was Ricardo who first emphasized that economics could be put into a form of pure theory and then used to abstract out the essential relationships. The historical school argued that it was impossible to understand how any real economy worked outside of the context of the society in which it is situated. The historical methodology emphasized exploring the actual shape of the economy and explaining how that shape came to be.

### Karl Marx

Marx's writings were to have a major impact on the twentieth century. Most people associate Marx with the concepts of socialism and communism, but in fact these ideas take up very little space in Marx's work. Instead, Marx spent most of his energy in describing and explaining the system of capitalism. What many are not aware of is that Marx based his descriptions of the capitalist system on the best economics of the time. Marx used the classical system as described by Ricardo as his basis of analysis. Marx took the labor theory of value and showed how the capitalist system was inherently unstable. The mechanics Marx used were not his own but rather taken from classical methodology. Needless to say, his works became very well known.

## The Marginal/Marshalian Methodology and the Rise of the Neoclassical

One of the major predictions of the classical economists was the eventual rise of a "stationary state" within the economic system. Malthus, Ricardo, and Mill all argued this would happen. Malthus and Ricardo both predicted that labor conditions would revert to subsistence. By the late 1800s, classical predictions were not materializing, labor conditions were improving, and there was no sign of an imminent end to economic growth. These failures in prediction led to rising criticisms of the classical paradigm, and new ideas began to appear. Some of the new thinkers were dissatisfied with the explanation of value arising out of the classical labor theory of value. Others wanted to explore more fully the nature of a self-directing economy. Whatever the differing reasons, several thinkers began to form the heart of a new methodology based on marginal thinking.

### The Beginnings of Marginal Thinking

The arrival of the marginal methodology was revolutionary, but the economists writing during the period were not aware of the revolutionary nature of the changes. Mark Blaug (1996), who is recognized as an expert on the history of economic thought, argued that the revolutionary nature of these changes was not really recognized until the next generation of economists. Yet what were the changes that happened, and who wrote about them?

Marginal thinking was a great breakthrough in economic theory. Simply described, marginal thinking suggests decisions are made on each consecutive choice rather than on whole groups of choices. Adam Smith posed a

puzzle called the water–diamond paradox that clarifies marginal thinking well. If someone were to offer you the choice of a bag of diamonds or a bottle of water, you would probably choose the diamonds. However, if you were in a hot desert and had not had water for several days, you would clearly choose the water. Why? In the first case, you are comparing the value of perhaps the first unit of diamonds to perhaps the thousandth unit of water. In the desert case, you are comparing the first unit of water with the first unit of diamonds. Marginal thinking suggests that decision makers decide to compare the extra units, not the total units. When comparing the first unit of diamonds to the thousandth unit of water, the diamonds are clearly more valuable. However, comparing the first unit of water to the first unit of diamonds quickly shows the value of water. Put in this manner, most people can see how they make decisions on the margin all the time, but clarifying this idea was revolutionary for the economic discipline.

Three major names are associated with the rise of marginal theory: Carl Menger, William Stanley Jevons, and Leon Walras. These three thinkers each operated independently in differing parts of Europe with amazingly different influences. However, even though they had little influence on each other and had few common influences on their own thinking, they simultaneously published major works elaborating on the ideas that were to become the marginal method.

Several major alterations in economic methodology arose during this period. The founders of the marginal methodology placed greater emphasis on demand versus the classical emphasis on supply. Such an emphasis on demand given the marginal perspective described above made sense. Why would someone choose the water over the diamonds? This is a demand question and is very different from what was being asked under the classical paradigm. The marginal movement argued that the value of an object was not constant as it was under the classical labor theory of value; rather, the marginal movement began using a subjective theory of valuation. One person may value the water more than the diamonds or vice versa, and who is to say one valuation is more accurate than the other? Marginal thinkers, particularly Walras, also emphasized a general equilibrium nature of the economy wherein all markets and decisions would balance out in a market system. Under a general equilibrium system, a small change in one area could affect all areas of the economy, but the flexible nature of the economy would sort out all of these changes automatically. Finally, the marginal founders also emphasized the process of rational maximizing behavior as the cornerstone of individual economic activity. Many of today's introductory economics students lose sleep over the ideas that originated at this time.

Such an emphasis on maximizing decisions based on subjective valuation led some to look more toward math to explore these differing individual choices. To some, math became a prevalent part of economic methodology, yet there was disagreement about the appropriateness of math as a tool for modeling the behavior of individuals. Menger, who is considered a founder of the Austrian school of economics, was strongly against the use of math. This antimathematical tendency is still prevalent in the Austrian school and its adherents. Jevons and Walras, on the other hand, used math as a tool to help explain their ideas. In the end, the work of Jevons, Menger, and Walras was incomplete. Their emphasis on the demand side of value was as flawed as the classical's emphasis solely on supply. The two sides would finally be brought together by economist Alfred Marshall.

## The Marshalian Scissors and the Neoclassical Method

Alfred Marshall was able to synthesize the works of many previous economists into a consolidated piece titled *Principles of Economics,* which was first published in 1890. The limited space of this chapter does not allow a detailed overview of the many contributors to neoclassical theory, but there were many. Each contributed to the ideas that were formulated into a single schema under Marshall.

Marshall was a great thinker, and his most important accomplishment was to bring together the newer works of the marginals and combine them with the important ideas from the classical school. Marshall took what was accomplished under the classical school and was able to build a partial theory of value based on costs. Marshall added the role of marginal decision making to firm profit maximization and explained a theory of supply. Marshall also took the theories of utility-maximizing consumers from the marginals and derived a theory of economic demand. Marshall then combined supply and demand into a single theory of market-driven price where both supply and demand work to determine price much like both blades of the scissors cut paper. Marshall's theory of markets has sometimes been referred to as the Marshalian scissors and is still a big part of what is learned in economics classes today.

With the works of Marshall, the foundations of neoclassical methodology were complete and contained the four components that would be major parts of economic methodology up until the 1990s. The first component was a vision of the economy as a great self-correcting machine similar to the Newtonian vision of physics common at the time. The second component was an emphasis on methodological individualism. The third was the use of deductive reasoning as a replacement for the experiments that were impossible to conduct, and the final component was an emphasis on internal consistency rather than external evidence. As described in the first section of the chapter, these were the major tools of methodology used by economists throughout the twentieth century.

## Objections to Neoclassical Economics

The rise of neoclassical economics was paralleled by a rise in thinkers critical of the methodology being used in

neoclassical economics. In economics, the methodology that is used by the majority of economics is often called the orthodox method, whereas the methodologies used by minority groups (usually critical of the majority) are called heterodox economics. Throughout the twentieth century, there were many heterodox critics of orthodox neoclassical methodology.

The historical school of thought was a heterodox group that was critical of the classical methodology and also actively criticized the neoclassical methodology for many of the same reasons. A controversy arose between the historical school and advocates of free market thinking. Part of the conflict revolved around the ability of a nation to plan an economy. Individuals arguing that such centralized planning could never work included Menger, whereas the historical school was led by a thinker named Gustav von Schmoller. The conflict between these differing groups was called the *Methodenstreit,* which is German for the "battle of the methods."

Another group of critics of neoclassical methodology were called the institutionalists. Thorstein Veblen, John R. Commons, and Wesley Mitchell were three founding institutionalists. Like the historical school, the institutionalists believed that economies had to be studied within a social context. The institutionalists also believed that the deductive method based on limited observations was flawed and that other methods were needed. Institutionalists criticized the neoclassical emphasis on the individual as the basis of study and argued in favor of a more socially driven vision of the economy. Institutionalists based their methodology on an evolutionary vision of the economy rather than a mechanistic view and argued that the theory of value created by the neoclassical methodology was rather simply a theory of price.

The Austrians were a third major group of critics of the neoclassical methodology. As mentioned above, Menger is considered both a founder of the Austrians and a major discoverer of marginal economics. Although the Austrian school is now considered a separate heterodox school, it has been considered only since the latter half of the twentieth century. Prior to the 1950s, the differences between the Austrians and neoclassical economics were small enough that a separating distinction was too minor to create a subclassification. The Austrian school moved away from the neoclassical school on a couple of points of methodology. The Austrians could agree with the neoclassical ideas that the individual should be the basic unit of economic study, and they could also agree that deduction was the best tool of reasoning available. The Austrians, however, began to diverge dramatically from the neoclassical perspective with respect to the usefulness of math in modeling individual behavior. As formal mathematical modeling of economics became more prevalent, so did the Austrian objections.

Although differing groups objected to neoclassical theory, this does not mean the differing groups agreed with each other. In fact, the founders of the Austrian school were some of the loudest critics of the historical school during the *Methodenstreit.* The Austrians and the institutionalists have also had sharp arguments over methodology. These differing schools still have proponents writing today and have influenced political policy in the past and present. The institutional school had substantial political influence from the beginning of the twentieth century until World War II. The Austrian school had a much greater influence during the second half of the twentieth century. In particular, Austrians such as Friedrich A. Hayak were highly influential on important political leaders such as Margaret Thatcher and Ronald Reagan.

## Keynes, the Neoclassical Synthesis, and Monetarism

The arrival of the Great Depression served as a major disruption to the neoclassical methodology. Neoclassical thinking took on a greater level of formality during the first three decades of the twentieth century, but little discussion was given about macroeconomic issues. Leon Walras had offered a partial explanation of a neoclassical vision of the macroeconomy in his theory of a general equilibrium model, yet a solid description of the model had never been given, and general equilibrium theories had been left unexplored until the 1930s. The Great Depression made this omission very clear and gave rise to the works of John Meynard Keynes and then eventually led to the inclusion of Keynes's ideas into a new version of the neoclassical methodology that used Keynes's insights and combined them in a Walrasian general equilibrium model. The 1970s saw a collapse of support for the neoclassical/Keynesian synthesis and a rise in the popularity of the monetarists.

### John Maynard Keynes

John Maynard Keynes was the son of John Neville Keynes, who was himself a well-known economist. Keynes was well trained in neoclassical methodology, and his works in macroeconomic theory were timely. Keynes looked back on the classical thinkers who spent most of their energy discussing the economy as a whole rather than economic activity on the individual level. Keynes had a daunting task to perform because he had to explain how a prolonged downturn in the economy could happen when most of neoclassical theory suggested that the economy would self-adjust. Keynes emphasized aggregate measures of economic activity to make a general theory of the macroeconomy. He began by analyzing how the goods created in the economy would be dedicated to different uses. They could be consumed, used for investment, used by the government, or traded to foreign nations.

With these expenditures explained, Keynes assumed that production in the system would follow what was desired by the differing groups who were acquiring the output. If the expenditures were maintained by the differing groups at a high level, then the economy would be

prosperous. However, there was nothing in Keynes's models to suggest that a high level of expenditures would automatically occur. In fact, Keynes envisioned situations wherein low levels of expenditures could remain for prolonged periods. In particular, Keynes suggested that during a crisis, firms would desire to invest significantly less, and because investment was a major part of expenditures, this could exacerbate the economic downturn. Simply put, a recession would cause firms to lower investment, which would further lower expenditures, which could lead to a bigger downturn. Keynes argued that this downward spiral could be stopped through government actions that increased government spending.

### The Neoclassical Synthesis

What Keynes wrote did not fit in nicely with the methodology of neoclassical economics. Keynes's work was not based in individual behavior as was neoclassical economics, did not emphasize a self-correcting mechanical nature to the economy, and lacked a mathematical formalism that was being strived for in neoclassical economics. However, the followers of the neoclassical methodology recognized the failure of their theory to reflect what was happening during the Great Depression, and the incorporation of the Keynesian vision into a neoclassical structure was quick.

The beginning of the synthesis occurred in 1937, when John Hicks wrote "Mr. Keynes and the 'Classics': A Suggested Interpretation," published in the journal *Econometrica.* The work was the beginning of a formalization of Keynesian ideas. The ideas were more formalized by other thinkers in the 1940s. The result of these works was a series of theories backed up with mathematical equations that were purported to represent the ideas presented in Keynes's work. The synthesis worked to place the ideas of Keynes within the context of the general equilibrium model first suggested by Walras. The resulting theories fit well within a neoclassical methodology, and Paul Samuelson (1948) popularized this method when he published it in his textbook that became one of the most popular economics texts available. During the 1950s and 1960s, economic analysis, as well as economic policy, was dominated by this neoclassical synthesis.

In Samuelson's text, much of the general information currently taught in undergraduate courses took form. The macroeconomy, based in large part on Keynesian expenditure analysis, formed the heart of the macroportions, and the market-based analysis of neoclassical economics formed the microportions. The formalization and apparent certainty of the neoclassical synthesis led to policy advice that seemed simple and useful. However, beginning in the 1970s, this certainty broke down under new economic circumstances. It should be noted that the models created in this synthesis failed to demonstrate how a continued period of economic downturn could happen, which was the main point of Keynes's work. There are some economists who have taken the Keynesian perspective in a very different direction. This group has created a fairly detailed description of a market system wherein continuing economic instability is explained, something that is missing in synthesis analysis. This group is called the post-Keynesians, and some of the better known economists from this group were Hyman Minsky and Joan Robinson.

### Monetarism

In the 1950s and 1960s, economic theory seemed so formalized that many began to think that governmental actions to correct economic imperfections were simple and easily accomplished. During the 1970s, the economy suffered a period of extended unemployment and inflation, which was something that could not be explained by the neoclassical synthesis models. This gave room in economic theory for a new group of thinkers led by Milton Friedman. Friedman and the monetarists argued that economic cycles were caused by changes in the levels of money in the economic system.

To the monetarists, all economic fluctuations were caused by changes in the monetary system. Friedman went so far as to try to demonstrate that the Great Depression was the result of misguided monetary policy and not the Keynesian explanation of dropping investment. With the decreasing popularity of the Keynesian position in the 1970s, the monetarists began to call into question the assumptions of the Keynesian positions writ large. The monetarists began to argue that the causes of the business cycles were changes in governmental policy and that if left alone the capitalist system was generally stable. The best way to maintain economic stability was to get the government out of the economy.

The monetarist ideas were different from the ideas of the neoclassical synthesis; however, the general methods still remained relatively the same. Starting with the foundations of the neoclassical school, there have been consistently four major components to economic method. The economy is seen as a mechanistic system with general equilibrium being the outcome. The basic unit of economic analysis is the individual, not groups. Deductive reasoning based on only limited observations is the best way to create economic theories, and internal consistency is to be valued over external evidence. Although new ideas and theories were to arise and fall, these basic tenets of methodology remained in economics for most of the twentieth century. This methodology is still dominant in undergraduate classes; however, the methods themselves are changing in the economics profession as a whole.

## Conclusion

Modern economic methodology is in a state of flux and is diverging from some of the main components of methodology that have been used for the past century.

Those previous components were first laid down during the neoclassical period of economics. However, modern economics seems to be moving away from them. Few modern economists use deductive reasoning as the basis for their work; instead, a much greater role has arisen for empirically testing theory using econometric tools. The rational individual as a cornerstone of theory is being replaced by a much more nuanced vision of the individual situated within society. External evidence that contradicts theories is given much more weight than ever, even if the evidence suggests that economic models may not be consistent. The mechanistic vision of markets always adjusting to equilibrium is also being questioned. Why all of this is happening now is difficult to answer. Perhaps the arrival of cheap and powerful computers is changing how economics is done. Perhaps it is the influence of critics of neoclassical methodology, or perhaps it is something else.

What is clear, however, is that the general concepts being taught in most undergraduate economics classes are the result of centuries of research and conflict. The origins of the micro- and macrodivisions of economics can be seen by looking back to the history of economic thought, and in that history you can find why methodology is taught as it is. This methodology is the result of an evolutionary process where differing ideas compete for acceptance. Whereas the undergraduate is taught economic methodology as a set of tools, the profession is always striving to alter and refine these tools to better understand the economy. In this process, the tools themselves change and methodology evolves. Adam Smith began asking questions important to his society at the time, and the tools he used fit those questions. Those tools have been altered by great thinkers such as Ricardo, Mill, and Marshall. Those tools have been shaped by important economists such as Malthus and Keynes. The methodology we receive today exists because of the works of these people, and the methodology our descendents receive will be altered by the economists writing today. Economics is a living discipline, and continued exploration will inevitably result in the metamorphosis of method. This chapter is intended as an introduction to the process; much more remains to be explored.

## References and Further Readings


Ariely, D. (2008). *Predictably irrational: The hidden forces that shape our decisions.* New York: HarperCollins.

Beinhocker, E. D. (2006). *The origin of wealth: Evolution, complexity, and the radical remaking of economics.* Boston: Harvard Business School Press.

Blaug, M. (1996). *Economic theory in retrospect.* New York: Cambridge University Press.

Colander, D. (2000). The death of neoclassical economics. *Journal of the History of Economic Thought, 22,* 127–143.

Colander, D. (2005). What economists teach and what economists do. *Journal of Economic Education, 36,* 249–260.

Colander, D. (2009). In praise of modern economics. *Eastern Economic Journal,* 35*,* 10–13.

Colander, D., Holt, R. P. F., & Rosser, J. B. (2007–2008). Live and dead issues in the methodology of economics. *Journal of Post-Keynesian Economics, 30,* 303–312.

Davis, J. B. (2007). The turn in economics and the turn in economic methodology. *Journal of Economic Methodology, 14,* 275–290.

Eklund, R. B., & Herbert, R. F. (1990). *A history of economic theory and method.* New York: McGraw-Hill.

Hicks, J. R. (1937). Mr. Keynes and the "classics": A suggested interpretation. *Econometrica, 5,* 147–159.

Keynes, J. M. (1936). *The general theory of employment, interest, and money.* New York: Harcourt, Brace & World.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Landreth, H., & Colander, D. C. (2002). *History of economic thought.* Boston: Houghton Mifflin.

Marshall, A. (1920). *Principles of economics.* New York: Macmillan. (Original work published 1890)

Mill, J. S. (1871). *Principles of political economy with some of their applications to social philosophy* (3rd ed.). London: Longmans, Green, Reader, and Dyer. (Original work published 1848)

Ricardo, D. (1996). *Principles of political economy and taxation.* Amherst, NY: Prometheus. (Original work published 1821)

Samuelson, P. A. (1948). *Economics, an introductory analysis.* New York: McGraw-Hill.

Schumpeter, J. A. (1960). *History of economic analysis.* New York: Oxford University Press.

Smith, A. (1965). *An inquiry into the nature and causes of the wealth of nations.* New York: New Modern Library. (Original work published 1776)

# 4

# Twentieth-Century Economic Methodology

Peter J. Boettke

*George Mason University*

The birth of the discipline of political economy is often dated to the eighteenth-century Scottish Enlightenment philosophers such as David Hume and Adam Smith. Of course, there were recognized antecedents dating back to Aristotle, the Spanish Schoolmen of Salamanca, and the French Physiocrats. Even the core idea of how private interest can be reconciled with public benefit through competition had a predecessor in Bernard Mandeville. But it was the Scottish philosophers who provided the foundation of classical political economy in the eighteenth century.

The systematic study of political economy begins with the recognition of two seemingly contradictory observations about commercial life. The first observation is that individuals pursue their self-interest and do so as effectively as they are capable of doing. The second observation is that commercial society exhibits a strong tendency to produce outcomes that enhance the public welfare in terms of material progress and betterment of the human condition more generally. Squaring these two observations is how the discipline was born.

## The Methodology of Economics and Political Economy: An Overview

Before we go further, I think it wise to stop and reflect on something unique about this disciplinary origin. Political economy and economics began with a reflection on an already existing set of practices in the world. It was, in this sense, in the quest to gain philosophical insight into the mystery of the mundane life around them that led these thinkers to study the economic system. In other words, a human practice was in operation that needed explanation. Economists did not invent economic life—whether as evidenced by the organization of the household, the harvesting of crops, the rise of manufacturing, or the free trade of goods and services across borders. Economic life happens, philosophers try to understand the manifestations of it—the changes in prices, the life and death of enterprises, the complexity of the division of labor, and the wealth and poverty of nations.

From the beginning of the discipline there have been debates concerning the methods used by thinkers to gain philosophic insight into these matters. One way to reconstruct Adam Smith's critique of the mercantilists is as a methodological critique of their understanding of the wealth of nations. Following the twentieth-century economist Fritz Machlup, this chapter will make a distinction between methods and methodology, where methods refer to the various techniques that economists employ in thinking about a problem and offering an explanation, and methodology refers to the philosophic study of those methods and their epistemological status. Methods of analysis have constantly evolved throughout the history of the discipline, and methodology has shifted as well with changes in epistemology. In other words, the positivism of the Vienna Circle placed criteria on what constitutes science that were different from the criteria that were understood during the age of British empiricism. As the criteria shift, so does the understanding of what is a good question to ask as well as what would be a good answer. Methods

of analysis are adopted if they help the explanation meet the currently fashionable criteria and discarded as relics of an older unscientific age if not.

Beginning in the late nineteenth and continuing throughout the twentieth century, the discipline of political economy was transformed into the science of economics as the methods employed by economists to study the economy more closely approximated the methods employed by those in the hard sciences, such as physics. Whether the methods developed in the sciences of nature were appropriate for the sciences of man was hotly contested throughout the twentieth century and continues to be the subject of intense debate into the twenty-first century. But it must be stated that most work-a-day economists do not see this as a debatable issue. Science is measurement, and the tools employed must satisfy that goal if science is to be done. The philosophical reflection on contemporary practice, let alone the entire enterprise of economics and political economy, is found in a specialized community of academics in philosophy, intellectual history, and economics who study the history of economic thought and methodology, as well as sometimes among the elderly of elite economists as they reflect back on their careers. In the discipline of economics proper, it is the very rare case (and professionally ill-advised) that a younger scholar will venture into the field of method and methodology.

But this does not mean that the methods of economics are stagnant either in the past century or today. No, they are constantly evolving as the problems that attract the attention of economists shift. However, the central disciplinary puzzle remains of explaining how through the self-interested behavior of individuals a social order can result that serves the public interest. The assessment of the truth value of this statement shifts with the times, as well as the normative assessment of economic exchange and the market economy. But every economist who has practiced the discipline since the eighteenth century would recognize the proposition that the market economy was self-regulating as central whether they agreed with it or not.

Joseph Schumpeter (1945), in his *History of Economic Analysis,* makes a distinction between "vision" and "analysis" and argues that "vision" is a necessary component of the advancement of scientific analysis. The simple reason is that "vision" is a pre-analytic cognitive act that provides the raw material for the scientist to analyze. As a mere matter of description of the way the human sciences operate, the economist must have a "vision" (a set of eyeglasses) that helps to clarify the questions that are to be raised. Visions are not neutral, however, with respect to the methods one uses to analyze a problem in the social world. Science may indeed be measurement, but the scientist has to know first what it is that must be measured and possess the measuring devises required for that task. Without either an idea of what to measure or the means for measuring, the intellectual enterprise can devolve quickly into nonsense rather than science.

Vision and analysis are both important parts of the narrative on the evolution of economic method and methodology in the twentieth and twenty-first centuries. The historical experience of economic disruption due to technological change, the devastation of war and depression, and the consequences of ideologically inspired revolutions also shaped twentieth-century economics. Just as the experience of the collapse of socialist ideological aspirations of the twentieth century shaped economic thought, the tragedy of less development, the fear of manmade global disaster such as irreversible climate change, the tensions of globalization, the fear of terrorism and religious fanaticism, demographic trends toward aging populations and the unsustainable public economic obligations that were made in the past, and the global financial crisis are in the process of shaping twenty-first-century economics. So we must be mindful of how visions frame the questions asked, how ideas of what constitutes science frame both what is considered a good question and good answer, how methods chosen will be a function of the question asked and the form an acceptable answer is expected to take, and, finally, how all of this is subject to change due to shifting philosophies of science, empirical puzzles that are thrown up in the world, and innovations in techniques of calibration that appear to permit measurement where it was seemingly impossible to get measurement before.

The toolkit of general competitive equilibrium, for example, that was developed in the late nineteenth century (Walras) and throughout the twentieth century (culminating in the Arrow-Hahn-Debreu model of the 1960s–1970s) sought to provide mathematical rigor to Adam Smith's "invisible hand" proposition. Smith's proposition came to embody in the mind of the economist a claim about the self-regulating nature of markets through relative price adjustments; the complex interdependency of economic life as evidenced by the division of labor, specialization, and exchange; and the efficiency of the market economy in production (least cost technologies employed) and exchange (gains from trade realized) through the guiding function of relative prices and the lure of pure profit and the penalty of loss. The incentives and information provided by clearly defined and enforceable private property rights; free movement of prices to reflect changing circumstances of tastes, technology, and resource availability; and profit and loss accounting, which induces entry of promising enterprises and weeds out failed enterprise, are enough to ensure that the market economy will satisfy the welfare criteria established by the theory of general competitive equilibrium. At least that is what elementary economics taught in the first chapters of Marshall's (1890/1972) *Principles of Economics,* as well as in the first chapters of Stiglitz's (1993) *Economics,* and for the most part every major textbook in between.

From at least the time of John Stuart Mill's (1843/1976) *Principles of Political Economy,* economists always have admitted that there were ample situations where the

"invisible hand" of the market would be hindered in its operation. The problem of monopoly was mentioned throughout the classical literature (though the source of monopoly was not seen in the natural tendencies of the market by many). The problem of common-pool resources was also mentioned, as were examples of what later would be termed externalities, asymmetric information with certain commodities, inequalities in distribution, economy-wide business fluctuations (theory of general glut or economic crisis), and public goods. The laissez-faire presumption that Mill laid out nevertheless had grounds for exception from the laissez-faire principle that were quite large. Many economic debates about method were in fact debates about how persuasive that case for the exception from the laissez-faire principle was. As analytical tools evolved, the answer to that question changed. Pigou had one answer, Coase had another, and Buchanan had yet another. To be clear, it is important to remember that if the laissez-faire principle stands, then the role of the economist is limited to that of a scholar and teacher, perhaps social critic, and the role of the government is mainly seen as that of a referee in the economic game. But if there are grounds for rejecting the laissez-faire principle, then the economists' role in society is potentially transformed into that of a policy engineer, and the government's role is transformed from a referee to an active player in the economic game. The method and methodology of economics are not invariant with respect to the policy aspirations of economists and political decision makers. It has been argued that the intended audience of Adam Smith's (1776/1976) *An Inquiry Into the Nature and Causes of the Wealth of Nations* was the enlightened statesman, and one could just as easily argue that this was true for the major works of Mill, Marshall, Pigou, and Keynes as well.

The theory of market failure developed by Paul Samuelson in the middle years of the twentieth century attempted to show under what precise conditions Smith's "invisible hand" proposition broke down. Ideas such as positive and negative externalities, free riders, excludability, nonrivalry, and so on became part of the everyday language of economists due to Samuelson's efforts both as a theoretical economist and as the leading textbook author for at least two generations of college students of economics. Samuelson's impact was in providing the latest reasons to doubt the veracity of Smith's proposition, and the form of argument in economics that he championed transformed the way economists must present their work for assessment among their peers. To eliminate ambiguity in argument, Samuelson argued, the rigor of mathematical formalism must replace the literary vagueness of an earlier less scientific age of economic analysis. The casualty of this transformation in method, Samuelson insisted, would only be the loose thinking of previous generations—loose thinking that produced an unfounded "faith" in laissez-faire and the invisible hand of the market economy. Samuelson spearheaded the neo-Keynesian synthesis in macroeconomics and the transformation of the welfare properties in microeconomics. Every area of economics, circa 1950 and 1960, was touched by Paul Samuelson. After Samuelson, the method and methodology would be unrecognizable to the previous century and a half of economic and political economy thinkers since Smith in a way that was not the case for, say, the history of the discipline from Smith to Frank Knight.

The Keynesian revolution and the development of macroeconomics in general offered an alternative vision that argued that in the context of the modern money-using economy, the classic link between private interest and public benefit had been severed, and thus the market could not be relied on to self-correct. Rather than self-correcting, the capitalist economy was said to be inherently unstable. Both the original Keynesian income expenditure model and the later neo-Keynesian IS-LM model were developed to demonstrate how once the link between savings and investment was broken, the classical vision of a self-regulating market economy that steered the self-interested behavior of individuals in such a direction that the public benefit was served could no longer be sustained. This model, not without challenges from thinkers such as Milton Friedman, dominated economic thinking in the post–World War II era.

Simultaneously with the lost faith in the central proposition of classical economics, economists also developed models that demonstrated that the market economy was prone not only to macroeconomic instability but also to monopolistic abuse and other microeconomic inefficiencies caused by various market imperfections. The model of general competitive equilibrium could no longer be said to mimic the outcomes of a free-market economy, but the model could serve as a tool of policy. The new approach to economics promised that government correctives would ensure that the welfare properties of the competitive equilibrium model would in fact be achieved even though the market economy could not achieve them when left to its own devices. The irony of this should not be lost. A model that was developed to represent what the market economy achieved without any central direction ("invisible hand") was transformed in the writings of economists such as Abba Lerner and Oskar Lange into a guiding tool for state direction of the economy (the visible hand of government planning). Economics was transformed from a discipline of philosophic reflection on the empirical reality of commercial life to a tool of social control by enlightened policy makers. Abba Lerner's (1944) book has the appropriate title, *The Economics of Control;* Samuelson, along with others, introduced linear programming into economics; William Baumol further developed the applications of operations research into economics; and some of the top minds in the field of economics, such as Leonid Hurwicz, would devote themselves to a field titled "mechanism design." In the 1940s to 1970s, the entire discipline of economics was transformed into a tool for social control, and

the methods and methodology of economics of that age fit that new purpose. Methods and methodology that did not fit the purpose of prediction and control were rejected as relics of a bygone era—a past era that was less scientific than the modern age.

In the mid-1970s, amid an economic reality of high unemployment and high inflation, the classical proposition concerning the "invisible hand" was resurrected with the work of Thomas Sargent and Robert Lucas and the new classical economics. Added to the lexicon of macroeconomics were rational expectations, time inconsistency, and the invariance proposition in policy design. The basic idea was that economists could no longer continue to model economic actors as completely passive actors that are to be manipulated by public policy decisions (as they were during the Keynesian hegemony) but had to model them as capable of anticipating the consequences of policies and therefore behaving in a way to put themselves in the best situation to take advantage of the policy change. This response, unfortunately, will potentially dampen the effectiveness of the proposed policy. A classic illustration of this was the Keynesian proposition concerning the trade-off between inflation and unemployment. The Keynesian consensus argued that as unemployment ticked up during a downturn, policy makers could stem this by engaging in inflationary policies. The inflation would drive down real wages, without affecting the nominal wage. In effect, workers will experience a wage cut, but due to "monetary illusion," they do not realize this, and so policy makers can keep unemployment in check through inflation. But this Phillips curve relationship breaks down if the workers recognize that their real wages are being cut and thus demand pay increases. Rather than inflationary monetary policy keeping unemployment in check, we get instead both inflation and unemployment rising. By the mid-1970s, the empirical reality of "stagflation" was the exact opposite of what was predicted by the Keynesian model of macroeconomics. Keynesian theory was empirically questionable, and theoretically incoherent was the judgment because it lacked microfoundations, and new classical macroeconomics filled the intellectual void.

At the same time, classical market theory reasserted itself against the market failure theories of the previous decade, and ideas such as the efficient market hypothesis (Fama), competition for the field (Demsetz), and contestable markets (Baumol) were added to the lexicon of microeconomics. Ronald Coase and James Buchanan pointed out that traditional Pigouvian welfare economics was logically either redundant (actors within the economy would bargain away conflicts themselves) or nonoperational (if private actors cannot bargain away the conflict, then under the same assumptions neither could public actors accomplish the policy goal). Coase and Buchanan spearheaded a revolution in economics to do comparative institutional analysis in law, politics, and the market. The conceptual framework of economics changed in the 1960s to 1970s, but many of those changes represented the resurrection of many of the themes found in the classical writings of David Hume, Adam Smith, David Ricardo, J. B. Say, and John Stuart Mill.

But it is important to stress for our present purposes that the changes of the 1960s and 1970s were not accompanied by a change in the methodology, so the methods were not so much transformed but applied consistently and persistently and into areas that previously were deemed out of bounds. In fact, one way to understand the new classical revolution was as a response to a dual intellectual inconsistency evident in the preceding economics literature: (a) a conflict between what was taught in microeconomics and macroeconomics in terms of core economic theory, thus requiring a search for microfoundations and (b) cutting short the story of market adjustment in microeconomic and macroeconomic narratives of imperfection and instability, such that when the economists opened up the analysis to account for agent learning and allowed for all the accommodating changes to take place in the market economy, the claims to imperfection and instability faded away. In other words, while there may be macroeconomic questions (e.g., inflation, unemployment, growth), there are only microeconomic answers, and those answers are provided through the examination of relative price effects and their impact on the behavior of individuals as they adjust to changing circumstances through time.

The development in the last quarter of the twentieth century of property rights economics, law and economics, public choice, the new learning in industrial organization, the economics of organization and new institutionalism, new economic history, entrepreneurial studies and market process theory, and new classical economics all represented efforts of one sort or another to analyze beliefs, behaviors, institutions, and situations that previously had been treated as either beyond the scope of analysis or as part of an unexamined framework. But again it is important to stress that while economic theory evolved and applications were found in new areas, the fundamental practice of economic methodology did not change as a result. In fact, the new methods were judged against the methodological conventions of formalism and positivism (at least the understanding of positivism among economists), and to the extent that the new methods failed to fit into those self-understandings of economic *science,* they would be dismissed as potentially interesting questions that were not operational. Until the set of questions being raised by new thinking in economics could be represented in a formal model subject to empirical test via sophisticated statistical analysis, they would have little impact on the practice of economists.

## The Fracturing of the Neoclassical Hegemony

I have argued that Paul Samuelson initiated the formalistic revolution in economics in the 1940s and 1950s. Samuelson's justification for this was simple—ambiguity

in thought emerges whenever we use the same words to mean different things or different words to mean the same thing, but by forcing economic arguments to be stated in a common formal language, assumptions would have to be made explicit (not hidden) and ambiguity would be avoided. In the 1950s, Milton Friedman also persuasively stated for economists that assumptions in theory construction did not really matter provided the construction was subject to empirical test. It is the submission to falsification that demarcates science from nonsense—an economist's rendering of logical positivism, instrumentalism, or whatever balled into an operational appeal for a simple formula of economic hypotheses subject to empirical test using statistical techniques. A philosophical statement of the positivist position with respect to the development of economics was actually made in the 1930s by T. W. Hutchison, but while recognized as a classic in economic methodology, the work did not persuade practicing economists. And it was not as if economists never made explicit methodological pronouncements—both descriptive and prescriptive prior to Hutchison. Lionel Robbins and Ludwig Mises defended the a priori and deductive logic nature of economic theory in the 1920s and 1930s. Mises, in particular, was adamant in his presentation of the a priori (purely deductive) nature of economic theory and built his argument on the earlier methodological work of N. Senior, J. N. Keynes, and C. Menger. Despite how Mises's statements have been interpreted by critics ever since, Mises did not claim originality for his position but argued instead that this was in fact the way that classical and neoclassical theorists of economics had in fact always done economics: deduction from self-evident axioms, combined with subsidiary empirical assumptions and aided by imaginary constructions (including the "method of contrast," where a world without change is constructed so we may understand the implications of change). In addition, Mises (following Weber) insisted on the positive nature of economic science against claims of ideological bias. Positive analysis prior to the philosophical development of logical positivism consisted of an argumentative strategy and was linked with Mises's consistent subjectivist stance. Treating ends as given and limiting analysis strictly to means-ends examination, Mises argued (as did Weber), would ensure the value-free nature of economics. Robbins (1932) picked up on this argument in the first edition of *An Essay on the Nature and Significance of Economic Science.* From a more continental philosophical tradition, Mises's student Alfred Schutz (1932/1967) made a similar argument in *The Phenomenology of the Social World.* However, the arguments of Mises and others that attempted to justify both methodological dualism (i.e., that economics was a science, but a science whose epistemic procedures were wholly different from those of the natural sciences) and the positive nature of economic theory proved to be ineffective in the wake of the empirical events of the 1930s and 1940s. The Great Depression and the grand ideological debates that were played out in World War II simply demanded an economics that was technical and analogous to physics—a form of social physics or better yet engineering. The discipline bent to this demand as young economists, motivated by the momentous events of the day, pursued advanced study of economics and went to work to solve social problems. Economics had to become a discipline capable of prediction and control and not endless disputes in social philosophy if progress was to be made and the shortcomings of the laissez-faire system were to be overcome through judicious public policy.

The change in the way economists do things that took place in the 1940s and 1950s was not led by philosophers of economics but a group of economic superstars who communicated to the rising generation how you are supposed to engage in the science. In this sense, the practice of economics in the second half of the twentieth century was dictated by Samuelson and Friedman (despite their disagreements), not by the Vienna Circle, Karl Popper, or Imre Lakatos. The methodological statements of Mark Blaug or Lawrence Boland or Bruce Caldwell in the 1970s and 1980s did not dictate practice in the discipline, just as the biting criticisms of Frank Knight or Ludwig Mises or Phil Mirowski (from the 1940s into the 2000s) have not curtailed the advance of the economists' self-understanding of the discipline as both formalistic and positivistic.

Both formalism and positivism came under intellectual assault in the 1960s to 1980s in the philosophy of science literature. Formalism resulted in unrealistic and sterile presentations of human life that missed as much as they captured, and positivism worked on an assumption that empirical tests were unambiguous. Without the empirical grounding provided by clean and unambiguous statistical tests, formal abstractions were prone to become free floating. Critiques of the modernist vision of science of an analytical form in the hands of Willard Quine (whether the falsifying result addresses the main hypothesis or the network of statements that led to the main hypothesis) or of the sociological variety in the hands of Thomas Kuhn and Michael Polanyi (paradigms and the notion of progress in science) or the continental form found in Richard Rorty (that all knowledge is contextual and framed by perspective) were embraced by various heterodox thinkers in economics, such as institutionalists, post-Keynesians, Marxists, and Austrian school economists. In addition, as formalism and positivism dominated practice in economics, there were always leading thinkers in the field who admitted the difficulties of carrying out the official methodology and questioned the current practice as failing to live up to the standards set or that the standards set were unrealistic. Ed Leamer's "Let's Take the Con Out of Econometrics" (1983) was one such critique of practice, as was D. McCloskey's (1998) *The Rhetoric of Economics.* There was, parallel to this, philosophers who challenged the economists' scientific pretensions, such as Alexander Rosenberg, who argued that economics was either mathematical politics or the science of diminishing returns, but

it was not an enterprise experiencing scientific progress. Dan Hausman has even described economics as an inexact and separate science. James Buchanan argued repeatedly that while economics is indeed a science, it is a philosophical science—which actually was a position staked out by R. G. Collingwood in the first decades of the twentieth century and was influential on Mises. Positions were stated by prominent figures, but they did not change practice or self-understanding.

During the 1980s and 1990s, survey articles on economic methodology were published in high-profile professional outlets such as *Journal of Economic Literature* and *Journal of Economic Perspectives,* and books on the subject were reviewed and discussed in a variety of traditional outlets. New journals were established such as *Economics and Philosophy* and the *Journal of Economic Methodology*. The critiques of traditional economic methodology led to a rise in heterodoxy, or at least a more self-confident and vocal heterodoxy. But ultimately, the actual methodological practice of elite economists changed little. McCloskey was a major advocate of change, but the criticisms offered in works such as *The Rhetoric of Economics* and *The Cult of Statistical Significance,* while widely read, did not have the force to change practice. What occurred instead was that attempts to justify practice by appeals to philosophy stopped among economists. Demarcation efforts were not a philosophical exercise but a complete embracing of scientific conventionalism. Economic science is what economists do, not what philosophers claim is scientific. And to do economics, one must think in terms of simplified models (parsimonious yet elegant mathematical representations) that are subject to sophisticated statistical tests. Critiques of the ability of statistical tests to answer fundamental questions (e.g., Greg Mankiw's critique of economic growth statistics) did not lead to a broadening in the notion of the evidentiary burden that must be met by contributions in the field but instead to renewed interest in finding better statistical instruments. Methods evolved, but the underlying methodology remained.

But the heterodox critique did not go completely unheeded. Questions concerning the behavioral foundations of economics led to a renewed appreciation of psychology and even neuroscience. Similarly, the institutionalist critique of economics led to a renewed examination of the legal-political-social nexus and its impact on economic life. Scholars such as John Davis have pointed to these intellectual developments among economists as evidence of a breakdown in the hegemony of the neoclassical mainstream. But one should be careful here because while psychological and institutional factors are now prominent in economic research, and the evolution of methods has led to a rise in laboratory experiments, computer simulations, and natural experiments that in a previous generation may have been viewed with suspicion, the form in which an argument must be stated in the top research journals to be considered "scientific" has not changed all that much from the days of Samuelson and Friedman.

## The Absorptive Capacity of Formalism

It is important to realize that I am not making a normative assessment of these developments (and lack of change) but instead providing a description of the intellectual landscape in economics from 1950 to today. Methods are constantly evolving but guided by a methodology that more or less has been fixed by scientific convention mid-twentieth century. Methodology not only determines to a large extent the questions that can legitimately be asked by a discipline but perhaps more important limits what would be considered a good answer to those questions. During this period, there have always been slightly out-of-sync economists who have been more or less nonconformists. Think of Nobel Prize winners such as F. A. Hayek, James Buchanan, Ronald Coase, Douglass North, Vernon Smith, and Thomas Schelling. Or broad-ranging economic thinkers such as Kenneth Boulding and Albert Hirschman. Or even recognized masters of the craft of thinking like an economist such as Armen Alchian. The economics profession during the twentieth and into the twenty-first centuries has had significant dissenters with respect to the prevailing consensus on method and public policy, but to dissent methodologically with respect to formalism and positivism was the quickest way to be utterly dismissed. And this was true even after developments in the philosophy of science literature questioned the modernist understanding of science. As McCloskey has repeatedly stressed to readers, economics is the most modernistic of the human sciences. And when the philosophy of science literature no longer justified those modernist ambitions, rather than rethink those ambitions, economists simply appealed to conventional practice. Economics is, in this understanding, simply what economists do. While previous generations of students were at least required to read Samuelson and Friedman on the methodology of economics during their first term in graduate school, the current generation of graduate students is expected to practice economics as they are taught without any serious study of the philosophical justification of the conventional methodology of economics.

Formalism has proven to be amazingly absorptive of heterodox ideas, especially after techniques and methods were developed that broke the taboo of multiple equilibria. Once the demand for models with determinate equilibrium results (i.e., single exit models) was relaxed, different paths could be explicated in models and so many heterodox ideas could be incorporated. One must remember that economists in the 1970s and 1980s struggled to gain acceptance for game theory, computer simulations, and laboratory experiments among economists. But once it was demonstrated that these methods could be used in a way consistent with the underlying methodology of model

and measure, they found wide adoption among economists for addressing questions that more traditional methods proved to be wanting. As Paul Krugman has pointed out in his discussions of the evolution of ideas related to economic geography and economic development, many nontraditional thinkers raised questions of increasing returns and location economies, but they lacked the tools to communicate those ideas in a way that economists could find useful. The usefulness criteria, I should point out, are provided by the methodological presumptions that were enforced. Useful, in other words, not as a tool of understanding but rather as a vehicle for forming testable hypotheses.

What is true for economic geography is also true for numerous other fields in economics, such as the study of politics, law, family, extended relationships, and norms. Many of the ideas being heralded as revolutionary are in fact the restatement of ideas held by an earlier generation of economists and political economists but previously deemed relics of an unscientific age of economics. Hume and Smith, for example, did not have a myopic view of humanity but instead believed in a behavioral model that included not only self-love but other regarding as well. The past 50 years of economic research and education have seen the placing of "old wine" of the classical school and early neoclassical writers into the "new bottles" of formalistic modeling and statistical testing empiricism.

The reports of the breakdown of orthodox hegemony by David Colander and John Davis have looked only at the method and policy dimensions, whereas the significant margin to look at that ultimately determines the character of economics is the underlying methodology. And on that margin, despite all the philosophical shifts, the basic justification of the enterprise of economics as a formalistic and positivistic science has remained unchanged. Unless an idea can be absorbed under this rubric, it will meet intellectual death. Methodology is the ultimate judge, jury, and executioner in economics, although it often lurks in the background unstated. As McCloskey has put it, when one's intellectual range is limited to M–N, when you get up close and look, it seems as if a wide range of topics are on the table and that all those around the table are fair and open-minded contributors to the enterprise, but when you step back, you realize that the intellectual span from M–N is quite narrow and misses the entire range of issues from A–L and O–Z. This is the fate of economics in its high modernist form, and little has changed in practice except that economists no longer appeal to high modernist philosophy to justify what they do.

## Where Is Economics Going?

With what I have said about the relationship between method and methodology firmly in mind, let us look at developments in economics as we entered the twenty-first century. First, during the last decade of the twentieth century, two empirical realities became of overriding concern to economists—the collapse of communism and the transition from socialism, and the failure of development planning and foreign aid programs to lift the less developed world into a position of greater freedom and prosperity. Second, as we prepare to enter the second decade of the twenty-first century, two other empirical realities became overriding concerns to economists—the tensions of globalization, threat of international terrorism, the financial crisis, and threat of worldwide depression.

One way to think about this is to envision the discourse in economics as following the shape of an hourglass. In the eighteenth and nineteenth centuries, economics was part of a larger discourse in political economy and moral philosophy. As the discipline self-identified with a more technical (less philosophical) and scientific approach to economic questions in the twentieth century, the scope narrowed. By the 1950s, the discipline of economics was narrowed to the midpoint on the hourglass. Since the 1950s, we have seen the broadening of the discipline again to take into account questions that once preoccupied the minds of the "worldly philosophers." By the turn of the twenty-first century, economists were once again tackling questions that could be recognized by the likes of Adam Smith and John Stuart Mill, let alone Max Weber. Indeed, the "worldly philosophy" seemed to be back en vogue. Amartya Sen tried to explain this shift in intellectual focus using the language of modern economics rather than imagery such as an hourglass. To Sen, there is a production possibility frontier for economic research, with economics as engineering on one axis and economics as philosophy on the other. During the twentieth century, economics moved toward a corner solution of economics as engineering, but during the last quarter of the century, economics began to move along the frontier away from the corner solution to once again pursue economics in a more philosophical manner. In Sen's writings, this shift relates to questions of ethics that must be raised for welfare judgments to be passed.

What I have suggested, however, is that while the questions have broadened once again, they have done so only to the extent that they can be restated in a form that conforms to the methodology that actually led to the narrowing of the hourglass (or the move along the frontier to the corner solution). I will leave to another time the question of whether this form constraint distorts the substantive content of the conversation. For now, the point I want to make is that the methodological constraint that produced the transformation of economics in the twentieth century (Samuelson-Friedman) is still binding on disciplinary discourse. Economics is a model and measure discipline; it is a discipline that advances through journal articles, not books, and it is a discipline whose scientific status while constantly questioned by outsiders is never questioned by those who occupy the commanding heights of the profession.

That much said, one would be blind not to see the important changes in research focus and methods of analysis that have taken place. In one sense, the empirical puzzles of the collapse of communism, the transition to capitalism, and the failure of development planning led to a renewed appreciation for the underlying institutional context of economic life. Economics in this sense is conceived of as a science whose subject is exchange and the institutions within which exchange takes place. On the other hand, the transplanting of institutions from one environment to another has proven to be a very difficult policy task, and both the transition to capitalism and the elimination of squalor and poverty in the Third World have proven to be more difficult than imagined. The contemporary empirical puzzles of the tensions emerging from globalization and the threat of international terrorism have focused economists' attention on "mental models," including ideology, religious beliefs, and cultural value systems in general. Directed by these pressing issues, economists are reexamining the cognitive foundations and behavioral assumptions of the discipline. In other words, in the upper echelons of the professional hierarchy, both the institutional and behavioral assumptions of conventional models have come under examination and a required modification.

As a question of public policy, many economists saw the link between the puzzles of the 1990s that resulted in examining the institutional assumptions and the puzzles of the 2000s that resulted in examining the behavioral assumptions. Solving the problems of transition and of Third World poverty has captured the imagination of leading economic thinkers such as Douglass North, Bob Lucas, Joe Stiglitz, Jeff Sachs, Esther Duflo, Abhijit Banerjee, Andrei Shleifer, and Bill Easterly. Many of these economists rejected the traditional neoclassical depiction of individual decision making and the market economy, as well as the free-market policy recommendations associated with the "Washington Consensus." Others argued for a more sophisticated understanding of the traditional model and offered a more nuanced defense of the basic message of the "Washington Consensus." These debates will continue. But the basic message from the economics of the 1990s and 2000s is that institutions matter, and while individuals respond to incentives, they are also prone to suffer delusions and other error-inducing cognitive limitations in those responses. Learning is context dependent. Behavioral economics, in particular, argues that individuals often are mistaken in their beliefs and expectations.

Actors are not perfectly rational, information is imperfect, markets are not atomistic, and resources are not always channeled to their highest valued use. While these admissions of imperfection open the discipline to new areas of research, those new areas can be pursued only via the conventional methodology. That is the dilemma of economics in the twenty-first century. Kenneth Boulding once remarked that the problem of economics (circa 1970) was that the discipline was asked to address twentieth-century problems (depression, war, cold war) with the mathematical tools of seventeenth-century physics (Newton). Hayek made similar remarks at the time—which should not be that surprising because both Boulding and Hayek were early adherents of general systems theory and respectively influenced in their thinking by Ludwig Bertalanffy and the idea of complex systems analysis. One could argue that an analogous critique could be offered to today's economics and does in fact get voiced in the discussions of complexity theory and economics. While models of social complexity and both agent-based and complex adaptive systems are not uncommon in the literature, they have not affected practice to the extent expected by the adherents of these models. In other words, the core theory of neoclassical general equilibrium theory remains the foundation of economic analysis. In a different context, Frank Hahn once described the criticism of the edifice of neoclassical theory as a bombardment of so many soap bubbles. In other words, the critiques are offered, but they ultimately bounce off.

Still, several method shifts in economics must be reckoned with in any discussion of economics and political economy in the twenty-first century. First, there has been the rebirth of political economy independent of the Marxist tradition. This goes back to the point raised by my hourglass metaphor or Sen's discussion of the movement along the production possibility frontier of economic thinking away from engineering and more toward philosophy. Positive political economy and constitutional political economy are intellectual developments in both the disciplines of economics and politics that have brought back the serious discussion among social scientists of the structure of government, the role of rules (formal and informal) in political and economic interactions, and the political-economic and legal-economic nexus. In one respect, post-1960 political economy can be accurately described as asking the fundamental questions of political theory from Thomas Hobbes, John Locke, David Hume, and U.S. founding fathers such as James Madison with the analytical tools of modern economics. When James Buchanan was awarded the Nobel Prize in 1986 for his development of public choice theory, political economy was well established again in the curriculum on economists and political scientists, and this has only continued in the decades since that award.

Second, there have been changes in the conceptual perspective of economists. There seems to be a willingness among economists to challenge the core ideas of rationality, self-interest, and equilibrium. But this willingness should be viewed with some suspicion because as I have argued, this new openness has been purchased by abandoning a commitment to substantive propositions in economics while steadfastly affirming the commitment to the form in which arguments must be made to be considered contributions to the economics. The criticisms associated with heterodox traditions of neoclassical methodology

have not won the day, and thus the enthusiasm one reads in Colander and Davis for the fracturing of the mainstream is overstated. Instead, the criticisms must be stated in a manner that conforms to those older Samuelson-Friedman notions of formalism and positivism. Roger Koppl has argued that the cutting edge of the mainstream (he refers to it as heterodox mainstream) is now occupied by work that employs the methods of bounded rationality, rule following, institutions, cognition, and evolution. And Koppl is certainly accurate in his description, but the argument that often accompanies this description of the current state of play in the discipline and the emerging alliance between various heterodox schools of thought such as post-Keynesian, old institutionalist, new institutionalism, complexity economics, Austrian economics, and post-Walrasian economics that will effectively challenge the prevailing orthodoxy is overstated. Instead, as I have stated earlier, the orthodoxy has tremendous absorptive capacity, and the evolution of methods of analysis such as evolutionary game theory has aided absorption. Heterodox arguments that can be restated in formal terms and tested using conventional statistical techniques can get a hearing among the professional elite, but those arguments that cannot quite be presented in that form (however interesting) will not get that same hearing, let alone influence economic research. This is one possible explanation as to why leading representatives of heterodox schools of thought are rarely published in the highest impact professional journals and are often unable to obtain teaching positions in the most prestigious departments. This is not an argument about discrimination and unfair barriers to entry in the field of economics. Economics is actually a very fluid discipline, and the culture at the top departments (e.g., University of Chicago) is notorious for the ruthless commitment to argument and not established status of individuals. But the judgment of what constitutes a good argument is not invariant with respect to the prevailing methodology. Model and measure rhetoric was used by Samuelson and Friedman to dismiss opponents, and the same can be seen today as the challenges of heterodoxy are absorbed into the orthodoxy—whether those challenges come from the lab, magnetic resonance imaging machines, computer simulations, history, anthropology, or philosophy. Still, there can be little doubt that the methods economists are employing in their work are evolving, and this evolution enables them to tackle many questions about the dynamic nature of economic life and the complex interdependencies that previous economic thinkers were unable to ask in a way that would produce acceptable answers as judged by the methodological strictures of formalism and positivism. Consider the work in this regard of the most influential economic thinkers in the 1990s and 2000s: Andrei Shleifer, Ed Glaeser, and Daron Acemoglu. These three have explored legal origins, the nature of regulation, and colonial heritage and the origins of democratic government. The questions are broad and the methods are creative, but the form in which the argument is stated is very conventional.

Third, there have been significant changes in the empirical techniques that economists employ in testing hypotheses. Developments in econometrics, such as nonparametric estimations, as well as instrumental variable approaches, have enabled economists to pursue empirical research on topics that previously had appeared elusive. In addition, there has been an acceptance of experiments across the board as providing not only useful but essential empirical information. Economists such as John List engage in natural experiments and field experiments. Of course, laboratory experiments have long been used by economists such as Vernon Smith to advance economic knowledge in the fundamental theory of choice, market theory, public goods, voting, and booms and busts. Smith's Nobel address is one of the most profound statements of the nature of rationality in economics, the context-dependent nature of choice, and the contingency of social order. Smith and his colleagues have studied trust relationships, cooperation in anonymity, conflict, and market efficiency. Also, developments in programming have enabled economists to do computer simulations that illuminate important economic ideas, such as the work on the interaction of zero information traders still able to generate market clearing. The focus on institutions has led to a renewed appreciation for the field of economic history among economists and political economists. The analytic narrative approach to political-economic history enables the rational choice theorist to combine the argumentative structure of economics with the compelling narratives of historical case studies (or comparative case studies). The bottom line: As the analytical methods of economics have broadened, they have been matched by new methods of empirical examination of the world around us. But note again that while the methods have evolved, the scientific aspirations of the intellectual enterprise have not—that aspiration is to provide a parsimonious model that generates testable hypotheses that are then subjected to empirical refutation.

One final change to the landscape of economics in the twenty-first century that is notable is the renewed interest in both the application of economics to unusual topics in everyday life and the popularization of economics among the public not as part of policy discourse but simply as a way of thinking about the world. This movement can be captured under the label "freakonomics" and is mainly associated with Steven Levitt. Levitt employs a natural experiment method to tackle everyday economics and make sense of statistical anomalies that are found in an examination of the data. Tyler Cowen's forays into "freakonomics" are more conceptual than Levitt's and attempt to walk his readers through the logic of choice, whereas Peter Leeson's work is focused more on explicating the mechanisms of social organization and explaining the operation of these mechanisms in unusual social environments.

As we finish the first decade of the twenty-first century, there should be little doubt that economics is a vibrant and diverse discipline. The methods of economics are constantly evolving as the technology of analysis changes. As I have discussed, in the twenty-first century, it has become commonplace for economists and political economists to tackle questions concerning the cognitive limitations of man and the institutional contingencies of exchange relationships. Koppl is right: The cutting edge of the profession is now occupied by researchers working on questions that were previously viewed as the domain of heterodox thinkers. But the narrative provided here disagrees with the assessment provided by Colander, Davis, and Koppl that a heterodox mainstream is emerging within economics that represents a fracturing of the core scientific enterprise of orthodox economics. When we look closer, what we see is that while the methods are evolving and the policy disputes are ongoing, the fundamental question of methodology and the conception of economics as a science are unchanging (and unchallenged). Economists are stuck in a world where the discipline attempts to mimic the methodology of the natural sciences. The new methods introduced (often imported from disciplines perceived as more scientific than economics to begin with) are always judged against this formalistic and positivistic standard. As long as this self-understanding and its corresponding standards of acceptance and rejection remain intact, then frameworks of analysis that focus disciplinary efforts on understanding rather than prediction will continue to be dismissed as unscientific. The challenges of the interpretative turn in the human sciences, as summarized by philosopher Richard Bernstein, are completely ignored. But today so are the admonitions by philosophers such as Alexander Rosenberg that economists must more faithfully follow the methodological prescriptions of positivism ignored. The formalistic and positivistic nature of economics is the product of scientific conventionalism, which actually proves to be a more elusive target in methodological disputes than explicit references to the philosophy of science.

Economics is what economists do, and what they do is build models and test those models against data sets with statistical tools. There are always exceptions to the rule, but the exception proves the point. Michael Polanyi once described how new contributions to science in general have to balance scientific plausibility, intrinsic interest of the community, and originality of the contribution. Economics is no different from physics in this regard. Conservative forces are weighed against revolutionary innovations in the practice of science to provide discipline so that wishful conjectures in truth seeking are channeled in a productive direction. Research efforts overlap, and the work of one scientist becomes the productive input into the scientific production process of another to form a dynamic orthodoxy. There has not been a revolutionary shock to the methodology of economics since the mid-twentieth century. There has been a broadening of topics and even the emergence of new and exciting methods in contemporary economics, but the basic notion of what it means to be doing scientific economics has not changed much since Paul Samuelson set the standard for theory and Milton Friedman explained what it meant to do positive economics. Methods of analysis are constantly changing and policy disputes are ongoing, but the underlying methodology of formalism and positivism has not been effectively challenged since it came to define the self-understanding of economics in the post–World War II period. So far, nothing in the twenty-first century practice of economics suggests that change to this self-understanding of scientific economics will come anytime soon.

## References and Further Readings


Bates, R., Greif, A., Levi, M., Rosenthal, J.-L., & Weingast, B. R. (1998). *Analytic narratives.* Princeton, NJ: Princeton University Press.

Bernstein, R. (1978). *The restructuring of social and political theory.* Philadelphia: University of Pennsylvania Press.

Bernstein, R. (1983). *Beyond objectivism and relativism.* Philadelphia: University of Pennsylvania Press.

Blaug, M. (1992). *The methodology of economics: Or how economists explain* (2nd ed.). Cambridge, UK: Cambridge University Press.

Boettke, P. (1997). Where did economics go wrong? *Critical Review, 11,* 11–65.

Boulding, K. (1948). Samuelson's foundations: The role of mathematics in economics. *Journal of Political Economy, 56,* 187–199.

Boulding, K. (1971). After Samuelson who needs Smith? *History of Political Economy, 3,* 225–237.

Buchanan, J. (1979). *What should economists do?* Indianapolis, IN: Liberty Fund.

Caldwell, B. (1982). *Beyond positivism: Economic methodology in the twentieth century.* London: Allen & Unwin.

Colander, D., Holt, S., & Rosser, B. (Eds.). (2005). *The changing face of economics.* Ann Arbor: University of Michigan Press.

Collingwood, R. G. (1926). Economics as a philosophical science. *International Journal of Ethics, 36,* 162–185.

Cowen, T. (2007). *Discover your inner economist.* New York: Dutton.

Davis, J. (2006). The turn in economics: Neoclassical dominance or mainstream pluralism? *Journal of Institutional Economics, 2,* 1–20.

Friedman, M. (1953). *Essays on positive economics.* Chicago: University of Chicago Press.

Hayek, F. A. (1948). *Individualism and economic order.* Chicago: University of Chicago Press.

Hayek, F. A. (1979). *The counter-revolution of science: Studies on the abuse of reason.* Indianapolis, IN: Liberty Press. (Original work published 1952)

Knight, F. (1956). *On the history and method of economics.* Chicago: University of Chicago Press.

Koppl, R. (2006). Austrian economics at the cutting edge. *Review of Austrian Economics, 19,* 231–241.

Lange, O. (1938). *On the economic theory of socialism.* Minneapolis: University of Minnesota Press.

Leamer, E. (1983). Let's take the con out of econometrics. *American Economic Review, 73,* 31–43.
Leeson, P. (2009). *The invisible hook: The hidden economics of pirates.* Princeton, NJ: Princeton University Press.
Lerner, A. (1944). *The economics of control.* New York: Macmillan.
Levitt, S., & Dubner, S. (2005). *Freakonomics.* New York: HarperCollins.
Marshall, A. (1972). *Principles of economics* (8th ed.). London: Macmillan. (Original work published 1890)
McCloskey, D. (1998). *The rhetoric of economics.* Madison: University of Wisconsin Press.
Mill, J. S. (1976). *Principles of political economy.* Fairfield, CT: Augustus M. Kelley. (Original work published 1848)
Mirowski, P. (1989). *More heat than light: Economics as social physics, physics as nature's economics.* Cambridge, UK: Cambridge University Press.
Mirowski, P. (2002). *Machine dreams: Economics becomes a cyborg science.* Cambridge, UK: Cambridge University Press.
Mises, L. (2007). *Human action: A treatise on economics.* Indianapolis, IN: Liberty Fund. (Original work published 1945)
O'Driscoll, G., & Rizzo, M. (1985). *The economics of time and ignorance.* Oxford, UK: Blackwell.
Polanyi, M. (1962). The republic of science. *Minerva, 1,* 54–74.
Robbins, L. (1932). *An essay on the nature and significance of economic science.* London: Macmillan.
Rosenberg, A. (1992). *Economics: Mathematical politics or science of diminishing returns.* Chicago: University of Chicago Press.
Samuelson, P. (1947). *Foundations of economic analysis.* Cambridge, MA: Harvard University Press.
Samuelson, P. (1948). *Economics.* New York: McGraw-Hill.
Schumpeter, J. (1945). *History of economic analysis.* New York: Oxford University Press.
Schutz, A. (1967). *The phenomenology of the social world.* Evanston, IL: Northwestern University Press. (Original work published 1932)
Smith, A. (1976). *An inquiry into the nature and causes of the wealth of nations.* Chicago: University of Chicago Press. (Original work published 1776)
Smith, V. (2007). *Rationality in economics.* New York: Cambridge University Press.
Stiglitz, J. (1993). *Economics.* New York: W. W. Norton.
Ziliak, S., & McCloskey, D. (2008). *The cult of statistical significance.* Ann Arbor: University of Michigan Press.

# 5

# ECONOMETRICS

PETER KENNEDY
*Simon Fraser University*

Several definitions of econometrics exist, a popular example being the following: "Econometrics is the study of the application of statistical methods to the analysis of economic phenomena." The variety of definitions is due to econometricians wearing many different hats. First and foremost, they are economists, capable of using economic theory to improve their empirical analyses of the problems they address. At times they are mathematicians, formulating economic theory in ways that make it appropriate for statistical testing. At times they are accountants, concerned with the problem of finding and collecting economic data and relating theoretical economic variables to observable ones. At times they are applied statisticians, spending hours with the computer trying to estimate economic relationships or predict economic events. And at times they are theoretical statisticians, applying their skills to the development of statistical techniques appropriate to the empirical problems characterizing the science of economics. It is to the last of these roles that the term *econometric theory* applies, and it is on this aspect of econometrics that most textbooks on the subject focus. This chapter is accordingly devoted to this "econometric theory" dimension of econometrics, discussing the empirical problems typical of economics and the statistical techniques used to overcome these problems.

There are two main differences between econometrics and statistics. The first is that econometricians believe that economic data reflect strategic behavior by the individuals and firms being observed, and so they employ models of human behavior to structure their data analyses. Statisticians are less willing to impose this kind of structure, mainly because doing so usually is not fully consistent with the data. Econometricians ignore such inconsistencies, so long as they are not gross, to enable them to address issues of interest. The second difference stems from the fact that most economic data come from the real world rather than from controlled experiments, forcing econometricians to develop special techniques to deal with the unique statistical problems that accompany such data. For example, when analyzing female wages, one needs to account for the fact that some women with children will appear in the labor market only if their wage is large enough to entice them away from being a homemaker; this means that a sample of female wage earners is not a random sample of potential female workers—other things equal, low-wage earners are underrepresented. Patching up statistical methods to deal with these kinds of problematic data has created a large battery of extremely sophisticated statistical techniques. In fact, econometricians are often accused of using sledgehammers to crack open peanuts while turning a blind eye to data deficiencies and the many questionable assumptions required for the successful application of these techniques.

Despite these and many other criticisms, Masten (2002) argues convincingly that econometricians have a crucial role to play in economics:

> In the main, empirical research is regarded as subordinate to theory. Theorists perform the difficult and innovative work of conceiving new and sometimes ingenious explanations for the world around us, leaving empiricists the relatively mundane task of gathering data and applying tools (supplied by theoretical econometricians) to support or reject hypotheses that emanate from the theory. To be sure, facts by themselves are worthless, "a mass of descriptive material waiting for a theory, or a fire," as Ronald Coase, in characteristic form, dismissed the contribution of the old-school institutionalists. But without diminishing in any way the creativity inherent in good theoretical work, it is worth remembering that theory without evidence is, in the end, just speculation. Two questions that

theory alone can never answer are: First, which of the logically possible explanations for observed phenomena is the most probable? And second, are the phenomena that constitute the object of our speculations important? (p. 428)

## Linear Regression

The three main applications of econometrics are *estimating* relationships describing economic behavior, *testing* hypotheses about economic behavior, and *predicting/forecasting* economic events. The main methodology employed for these purposes is regression analysis, the essence of which is that we are interested in estimating a relationship such as

$$\text{Wage} = \alpha + \beta*\text{Education} + \delta*\text{Male} + \varepsilon.$$

This is interpreted as saying that an individual's wage is determined as a linear function of his or her years of education, gender (a "dummy" variable equal to 1 for males and 0 for females), and a random error term ε. The unknown parameters in this relationship, α, β, and δ, are what we wish to estimate; econometricians typically use Greek letters for unknown parameters. Special values of these parameters, such as that δ is equal to zero (corresponding to no discrimination on the basis of gender), are what we wish to test. Finally, if we knew the gender and years of education of a new individual, we could use this estimated equation to forecast this individual's wage.

Students studying econometrics need to become comfortable with several facets of regression.

The equation above is called the *specification;* the specification denotes the set of variables appearing in that relationship and the functional form of the relationship—in this case, a linear functional form. The wage variable is called the *dependent variable* or the *regressand.* The education and male variables are called *independent variables, explanatory variables,* or *regressors.* The unknown parameter α (unknown parameters are usually denoted by Greek letters) is called the *intercept* or *constant* term. The β and δ parameters are called *slope coefficients.* When there is more than one explanatory variable, the regression is called a *multivariate* regression. We speak of running a regression of the dependent variable on the set of independent variables; unless explicitly stated to the contrary, this includes an intercept.

Although this equation is written in a way that suggests a causal relationship, it is important never to forget that regression analysis can only assess the strength and direction of a quantitative relationship involving these variables. *Any conclusions regarding causality must come from common sense and economic theory.*

The error term is the *random,* or *stochastic,* element of the equation. It is added on for several reasons. First, a million extra variables influencing wage have been omitted from this equation; the error term reflects the sum of the effects of all these omitted variables. Second, there usually is some measurement error associated with the dependent variable. Third, the functional form of the relationship is probably not linear. And fourth, human behavior typically has a purely random component—faced with the same set of circumstances, a person will not always do the same thing. The error term is assumed to have zero expected value; any nonzero expected element of the error is absorbed by the intercept term.

The nonstochastic part of this relationship—namely, the equation without the error term—is called the *conditional expectation* of wage. What this means is that given values for the explanatory variables, the expected value of wage is given by this equation.

The slope parameter of an explanatory variable is the amount by which the dependent variable changes when that explanatory variable increases by one unit, *holding all other explanatory variables constant.* The terminology *ceteris paribus* is often used to refer to holding all other variables constant.

Estimates of the coefficients α, β, and δ are usually called $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\delta}$. The forecasted wage for a man with 12 years of education is $\hat{w} = \hat{\alpha} + 12\hat{\beta} + \hat{\delta}$. The forecasted wage for a woman with 15 years of education is $\hat{w} = \hat{\alpha} + 15\hat{\beta}$. In words, we set the unknown error term equal to its expected value (zero) and plug our explanatory variable values into the estimated specification to predict/forecast the dependent variable.

The error in predicting the *i*th value of the dependent variable, $w_i - \hat{w}_i$, is called the *i*th *residual* and is denoted $e_i$ or $\hat{\varepsilon}_i$ because it is an estimate of the *i*th error term, $\varepsilon_i$. The popular ordinary least squares (OLS) estimates of the unknown parameters are estimates that result from finding values of the parameters that minimize the sum of squared residuals (SSR). The terminology sum of squared errors is often used in its place and so is often denoted SSE.

Because it minimizes the sum of squared residuals, the OLS estimator automatically maximizes the "fit" of the estimated equation, as measured by $R^2$, the fraction of the "variation" in the dependent variable explained linearly by variation in the explanatory variables. Adding an explanatory variable helps the computer minimize the SSE, even if this explanatory variable is irrelevant; the *adjusted* $R^2$ (written $\bar{R}^2$) corrects for this. $\bar{R}^2$ is only one of many criteria relevant to the choice of specification, primary among which is economic reasoning.

Thanks to the power of modern computers, calculating the OLS estimator, as well as most of the other estimators that econometricians employ, can be done very quickly and easily with econometric software. The more prominent of these are EVIEWS, LIMDEP, PC-GIVE, RATS, SAS, STATA, and TSP. An illustration of regression output produced by econometric software appears later.

## Sampling Distributions

### Why Is OLS So Popular?

It is extremely important to realize that the fact that OLS minimizes the sum of squared residuals and so

"fits" the data best (has the highest $R^2$) is not the reason why OLS estimates are so popular. OLS is popular because in a classic estimating problem (the classical linear regression [CLR] model), it has a high probability of generating an estimate "close" to the true value of the parameter being estimated. But if the estimating problem at hand does not satisfy the assumptions of the CLR model, OLS could be a very bad estimator. The study of econometrics revolves around how to generate a "good" estimate in a given estimating situation. All this is formalized in the sampling distribution concept, the foundation on which the logic of classical statistics rests. If you get this logic, statistics/econometrics makes sense! Here are the main things that need to be understood about sampling distributions.

Loosely speaking, a sampling distribution tells us the relative frequency with which we would obtain different values of a statistic (such as $\hat{\beta}$, an estimate of the slope of education in the specification given earlier) if we were to calculate that statistic many times over, each time using data embodying a new set of randomly drawn error terms. The important implication here is that the value of the statistic that we actually obtained can be viewed as a single random drawing out of this hypothetical sampling distribution; it is determined by the unknown error terms in the actual data we used. Here is an example. Suppose each student in a class of size 200 interviews 10 randomly chosen students and averages their 10 ages to produce an estimate of the average age of all students on campus. Would these numbers all be the same? No. We could take these 200 numbers and use them to create a histogram. This histogram would picture the sampling distribution of the sample average statistic for sample size 10. So if we had only one sample (of size 10), it could be viewed as a random draw out of this distribution.

To illustrate this, suppose that the variable $y$ (wage?) is a linear function of the variable $x$ (education?) plus an error term $\varepsilon$ with mean zero, so that we have $y = \beta x + \varepsilon$, where for simplicity we have omitted the intercept. Given 25 observations on $y$ and $x$, there are several possible ways of using these data to estimate the unknown parameter $\beta$. One of the simplest is $\beta^* = \Sigma y/\Sigma x = \beta + \Sigma\varepsilon/\Sigma x$. This shows that the value for $\beta^*$ is equal to $\beta$ plus (or minus) an amount that depends on the random errors we drew when we obtained our data. The errors have mean zero, so summing them should involve a lot of canceling out as positive errors are offset by negative errors. This suggests that $\beta^*$ should be a pretty good estimate of $\beta$ because $\Sigma\varepsilon/\Sigma x$ will be small, resulting in $\beta^*$ being close to $\beta$. But $\Sigma\varepsilon/\Sigma x$ will not be zero except by a fluke; sometimes it will be a positive number and sometimes a negative number, depending on the particular unknown error terms inherent in our data. It will probably be a small positive number or a small negative number, but if we had a peculiar draw of error terms, it could, with low probability, turn out to be a large positive or a large negative number. It is these possibilities, with different probabilities of occurring, that give rise to the sampling distribution of estimating formula $\beta^*$.

## How Do We Know What the Sampling Distribution Looks Like?

The computer output from an estimation procedure (such as OLS) provides a coefficient estimate $\hat{\beta}$ along with an estimate of its standard error. $\hat{\beta}$ is a single draw from its sampling distribution, so this does not help much in identifying what the sampling distribution looks like. But $\hat{\beta}$'s standard error is an estimate of the standard error of the sampling distribution, so this does provide information about $\hat{\beta}$'s sampling distribution, particularly about its spread. Beyond this, there are three ways of discovering what the sampling distribution of a statistic looks like—in particular, what is its mean. One has to do some algebra to derive it theoretically. This can be done in simple cases, but in most cases, the algebra is too difficult. In these difficult cases, two alternative methods are available. One is to use *asymptotic* (assuming the sample size is extremely large) algebra, which simplifies the algebra immensely but forces us to assume that results that are true for very large sample sizes are at least approximately true for our actual sample size. The other is to use a Monte Carlo study, in which a computer simulation is used to estimate the sampling distribution.

## Why Are Sampling Distributions So Important?

In short, the answer is because any statistic we calculate can be considered a random draw from that statistic's sampling distribution. This in turn has two extremely important consequences, one for statistics being used to estimate an unknown parameter and the other for statistics being used to test null hypotheses.

Suppose we are trying to choose between two estimating formulas, $\beta^* = \Sigma y/\Sigma x$ and $\beta^{**} = \Sigma xy/\Sigma x^2$, to estimate an unknown parameter $\beta$. Our choice is really the following: Would we prefer to estimate $\beta$ by drawing randomly a number out of $\beta^*$'s sampling distribution or by drawing randomly a number out of $\beta^{**}$'s sampling distribution? So to choose our estimator, we look for an estimating formula that has the "best-looking" sampling distribution.

## What Are the Characteristics of a "Good-Looking" Sampling Distribution?

One characteristic is *unbiasedness:* The mean of the sampling distribution equals the unknown parameter value $\beta$. A second characteristic is *efficiency:* The variance of the sampling distribution is small. A third characteristic is minimum *mean square error* (MSE). MSE is the expected magnitude of the squared distance between an estimate and the parameter it is estimating. A *best unbiased* estimator is an unbiased estimator that has variance smaller than the variance of any other unbiased estimator. A minimum MSE estimator is used whenever it is not possible to find an unbiased estimator with a small variance; it accepts some bias to reduce variance. This trade-off comes from deriving that MSE is equal to the sum of variance and squared bias.

An estimator that has a good-looking sampling distribution for one problem may have a bad-looking sampling distribution for another problem. In an earlier example, $\beta^* = \Sigma y/\Sigma x$ had a good-looking sampling distribution. But if in that example the intercept were not zero, this estimator would have a bad-looking sampling distribution. (Redo this example with an intercept $\alpha$, obtaining $\beta^* = \Sigma y/\Sigma x = \beta + N\alpha/\Sigma x + \Sigma\varepsilon/\Sigma x$, where $N$ is the sample size. Now $\beta^*$ will be close to $\beta$ only if $\alpha$ is close to zero!) A consequence of this is that a lot of what econometric theorists do can be characterized as follows: For a new estimating problem, find the estimating formula with the best-looking sampling distribution.

The second major use of the sampling distribution concept is in the context of hypothesis testing. Suppose we wish to test that $\beta = 1$, for example, and to that end, we calculate a test statistic $q$. Because $q$ is a statistic (i.e., it is a number calculated by putting the data into a formula), it has a sampling distribution. We ask ourselves the following: What would this sampling distribution look like if the null hypothesis were true? A lot of what econometric theorists do can be characterized as follows: For a testing problem, find a test statistic that, if the null hypothesis is true, has a sampling distribution described in one of the tables found at the back of statistics books.

The rationale behind hypothesis testing is intimately connected to the sampling distribution concept. A test statistic can be viewed as having been obtained by drawing randomly a single number out of that statistic's sampling distribution. Obtaining a number from the tail of the appropriate sampling distribution tabulated at the back of a statistics book (which assumes the null is true) poses the following dilemma: *Did we obtain this number by chance, or did it in fact come from a different sampling distribution, one that characterizes this test statistic when the null hypothesis is false?* The hypothesis testing procedure in common use results from setting up a rule to create an answer to this question.

### What Is This Rule?

If the number comes from the $\alpha$% tail of the null-is-true sampling distribution, conclude that the number came from a null-is-false sampling distribution and so reject the null hypothesis. In this case, $\alpha$% is the arbitrarily chosen Type I error; it is most commonly (without any good reason!) chosen to be 5%.

## The Classical Linear Regression Model

Introductory econometrics texts revolve around the CLR model, a set of assumptions about how the data were generated. If the data have been generated according to this model, the OLS estimator has very desirable sampling distribution properties, making it the first choice for estimation. Violation of one or more of the assumptions of the CLR model means that econometric theorists need to figure out what implications this has for the desirability of the OLS estimator and may lead to the choice of an alternative estimator. Chapters in textbooks typically examine violation of these assumptions one by one. There are five basic assumptions in the CLR model.

1. The first of these assumptions is that the linear specification is correct and that we have the right set of explanatory variables. Consider the specification used earlier:

$$\text{Wage} = \alpha + \beta^*\text{Education} + \delta^*\,\text{Male} + \varepsilon$$
or
$$w = \alpha + \beta\text{ED} + \delta\text{Male} + \varepsilon.$$

In this specification, wage is a linear function of years of education and gender, plus an error term. It also specifies that wage is determined only by education and gender, so that the influence of a million other variables is adequately captured by the error term. Probably some of these million other variables, such as experience and ability, are sufficiently important that, if possible, they should explicitly be included as explanatory variables in the specification.

The linearity assumption is not as restrictive as it appears. Here are some functional forms that reduce to linearity for estimation purposes.

a. *Polynomial:* $w = \alpha + \beta\text{ED} + \gamma\text{ED}^2 + \delta\text{Male} + \varepsilon$.

Here the square of years of education is included as an explanatory variable to allow the influence of education to die off as years of education becomes bigger and bigger. Note that this implies that the interpretation of $\beta$ changes. Although the functional form is quadratic, wage is a linear function of ED, $\text{ED}^2$, and Male, so that we can estimate via a linear regression of $w$ on ED, $\text{ED}^2$, and Male. What this means is that we create a new explanatory variable called $\text{ED}^2$, which has as observations the squares of the ED observations.

b. *Log-linear* or *semi-logarithmic:*
$\ln w = \alpha + \beta\text{ED} + \delta\text{Male} + \varepsilon$.

Here the dependent variable is the natural log of $w$, allowing the specification to represent a situation in which it is the percentage change in the dependent variable that is determined by changes in the explanatory variable. ($\beta$ is the percentage change in $w$ due to a unit change in ED.) This is in fact the most common way of modeling wage equations. Although the functional form is semi-logarithmic, $\ln wage$ is a linear function of the explanatory variables so that we can estimate via a linear regression of $\ln w$ on ED and Male.

c. *Log-log* or *double-logarithmic:*
$\ln w = \alpha + \beta\ \ln\text{ED} + \delta\text{Male} + \varepsilon$.

Here both the dependent variable and one or more of the explanatory variables are in log form. In this case, β is an elasticity—the percentage change in the dependent variable per percentage change in the explanatory variable. Estimation is via a linear regression of ln$w$ on lnED and Male. The classic example of this functional form is the Cobb-Douglas production function, in which output $y$ is a function of capital $K$ and labor $L$:

$$y = AK^{\alpha}L^{\beta}\varepsilon,$$

which upon taking logs becomes

$$\ln y = \ln A + \alpha \ln K + \beta \ln L + \ln \varepsilon.$$

In this functional form, α and β are elasticities, and α + β is the returns-to-scale parameter. A general rule for deciding when to use logged variables is to ask whether percentage changes or absolute changes are relevant for the context of your problem. For the wage specification, for example, people usually think in terms of percentage wage changes and the absolute number of years of education. In general, wages, income, price indices, and population figures are logged, and age, years of education, and rates of change such as interest rates are not logged.

2. The second assumption of the CLR model is that the expected value of the error is zero. In one respect, this is an innocuous assumption because the intercept in the specification gathers together the average influence of the million omitted explanatory variables, allowing the error term to have zero expected value. It is tempting to interpret the intercept in a linear functional form as the value of the dependent variable when all the explanatory variables are zero. This is a misinterpretation of the role of the intercept. The intention of all functional forms, linear or nonlinear, is that they approximate the "true" unknown functional form throughout the range of the data. The intercept merely serves to enhance this approximation.

3. The third assumption of the CLR model is that the error term variances are the same for all observations and that the error terms are all independent of one another. Violation of this assumption takes two common forms.

a. *Heteroskedasticity:* The variance of the error terms is not the same for all observations. A classic example of this is when the dependent variable is an average across all households in a town, with each town contributing one observation. Because the towns do not all have the same population, the averages will have different variances.

b. *Autocorrelated errors:* The error term for one observation is influenced by the error term for another observation. The most common case is in time-series data when the error in period $t$ is affected by the magnitude of the error in the preceding period $t - 1$. If an earthquake affects the dependent variable via a large negative error in time period $t$, the effect of the earthquake probably will not be fully overcome by the following period, so it is likely that the error in period $t + 1$ will also be negative.

4. The fourth assumption of the CLR model is that the explanatory variables are uncorrelated with the error. A classic example of this is simultaneity, a common economic phenomenon. Suppose we are estimating a demand function, regressing quantity on price. We know that quantity is determined by the intersection of the supply and demand curves. If the error term in the demand function bumps up, it shifts the demand curve, which through the simultaneous interaction with the supply curve changes price. Consequently, the error in the demand curve is correlated with price, the explanatory variable in the demand curve. Because when explaining changes in the dependent variable the OLS procedure gives as much credit as possible to the explanatory variables and as little credit as possible to the errors, the OLS procedure in this case lumps together the influence of the explanatory variable price and the influence of the error term and so creates a misleading (i.e., biased) estimate of the slope coefficient on price.

5. The fifth assumption of the CLR model is that there is no exact linear relationship among the explanatory variables. A classic example is the dummy variable trap. In our earlier example, we had a variable Male for gender, defined as 1 for males and 0 for females. Suppose we added an extra dummy variable Female, defined as 1 for females and 0 for males. Then Male plus Female equals 1, exactly equal to the intercept (the implicit 1 multiplying α). In this case, if you run a regression, the computer will rebel, complaining that you have attempted an impossible calculation such as trying to divide by zero. (More likely it will say "near singular matrix.") This problem is called *perfect multicollinearity;* it is impossible to run the regression. The case of an approximate linear relationship (which allows the regression to be run) among the explanatory variables is referred to as the *multicollinearity problem.* In this case, because the explanatory variables move together a lot, it is very difficult for the computer to figure out which explanatory variable is responsible for changes in the dependent variable. This causes the OLS estimates of the slope parameters to be unreliable: The variance of the OLS estimate is large.

6. An extra assumption that is sometimes added to the CLR model is that the errors are distributed normally. This creates the CNLR model, the classical *normal* linear regression model. The advantage of this assumption is that the normality of the errors allows use of the tables at the back of the book for hypothesis testing. Without normally distributed errors, the numbers in these tables are correct only in large samples. Fortunately, in most cases in samples of very modest size, these tables provide very good approximations to the numbers needed to perform hypothesis tests.

## Violating the CLR Model Assumptions

If the assumptions of the CLR model are met, the OLS estimator is unbiased and has the smallest variance among all linear estimators (a linear estimator is a linear function of the errors), and so it is called the *best linear unbiased estimator* (BLUE). The proof of this is called the *Gauss-Markov theorem.* But what happens to these desirable properties of the OLS estimator if the CLR model does not hold? Let us look at each of the CLR model assumptions in turn.

1. *Specification error.* The classic specification error dealt with in textbooks is omission of an important explanatory variable. In our wage example, suppose a measure of ability is omitted. Ability is likely correlated with education because people with more ability are likely to take more years of education. Because of this correlation, education will get credit for some of the influence of ability, biasing the education coefficient estimate. This could be a good thing if the estimated equation is to be used for forecasting because it compensates for the missing ability variable. But if the purpose of estimation is to produce a "good" coefficient estimate on education, the bias will not be welcome. Finding a wrong sign on a coefficient estimate is a warning that something is wrong with the specification. Because the specification is unknown, doing applied econometrics is much more difficult than doing econometric theory.

2. *Nonzero expected error.* If the expected value of the error term is a nonzero constant, no problem is created because its constant component is absorbed into the intercept.

3. *Heteroskedasticity or autorcorrelated errors.* There are two major consequences of errors that are heteroskedastic or autocorrelated. First, although OLS remains unbiased, it is not as efficient as an alternative estimator, the *generalized least squares* (GLS) estimator. This estimator minimizes a weighted sum of squared errors, where, for example, a lighter weight is put on errors with bigger variances because these observations are less reliable. But to some, a more important implication of violating this assumption is that the standard errors of the parameter estimates are biased, fouling up hypothesis testing. Whenever heteroskedasticity or autocorrelated errors are suspected, econometricians use OLS but employ "robust" standard error estimates that avoid this bias.

4. *Error correlated with an explanatory variable.* An instrumental variable (IV) estimator is usually employed to circumvent the bias caused by violation of this assumption. An IV is uncorrelated with the error but correlated, preferably highly so, with the delinquent explanatory variable. In our demand function example above, suppose that weather affects supply but not demand. Clearly, the weather is not correlated with the demand function error. And because it affects supply, it shifts the supply curve and so affects price, causing weather and price to be correlated. The part of price that is explained by weather (call it $\hat{p}$) is not correlated with the error, so by regressing quantity on $\hat{p}$, an (asymptotically) unbiased estimate of the price coefficient can be obtained. A drawback of the IV estimator is that its variance is bigger, sometimes dramatically so, than the variance of the OLS estimator; because of this, OLS may be better on the MSE criterion.

5. *Multicollinearity.* No bias is created, and OLS continues to be BLUE. The only problem is that the standard errors of the OLS estimator are large. There are only two solutions. First, do nothing because although the standard errors are large, they may not be so large as to compromise the purpose of the analysis. Second, find and use more information. More information means smaller standard errors. A classic example is to drop one of the explanatory variables that is highly correlated with another, adding the information that its coefficient is zero or nearly so. This eliminates the collinearity. It creates bias (deliberately!) because what you had believed was a relevant explanatory variable is omitted, but by eliminating the collinearity, standard errors shrink. On the MSE criterion, the resulting estimator may be superior.

## Beyond the Basics

What has been described above represents the basics of econometrics:

- the sampling distribution concept, upon which the logic of traditional econometrics rests;
- the CLR model, the workhorse of regression analysis; and
- the major violations of the CLR assumptions, the structure around which econometrics textbooks are built.

There is of course far more to econometrics. To provide a flavor of this, several major topics addressed in econometrics courses are briefly discussed.

### Nonlinear Regressions

Although as noted earlier, a lot of nonlinear functional forms can be estimated as linear regressions, many cannot. Here are some examples.

*Qualitative dependent variable.* If the dependent variable is a dummy variable—for example, taking the value 1 if an individual takes public transportation to work and 0 if not—a logit or probit model is used to estimate. These models have functional forms that produce dependent variable values lying between 0 and 1, with the estimated dependent variable values interpreted as the probability that the individual in question will take public transport. Multinomial logit/probit

models are extensions of this, with the dependent variable having more than two qualitative outcomes—for example, taking the bus to work, taking the subway to work, or using private transportation. If the dependent variable is qualitative with a specific order, such as bond ratings being AAA, AA, A, or B, an ordered logit or ordered probit can be used for estimation.

*Limited dependent variable.* If the dependent variable is limited in some way—for example, the number of tickets sold for a ballgame cannot exceed the capacity of the stadium—a tobit model may be used to estimate the demand function for tickets. Sample selection models are more sophisticated versions of this. For example, a female may enter the labor market only if the wage she is offered exceeds her reservation wage. The model thus contains two equations, one determining the reservation wage and the other determining the actual wage; estimation is complicated by having to recognize that only certain kinds of women make it into the sample of working females.

*Count data.* If the dependent variable is an integer representing the number of times something occurs, such as the number of cigarettes smoked in a day, a Poisson model may be used for estimation. The existence of nonsmokers complicates this because they all smoke zero cigarettes every day. The hurdle Poisson model adjusts for this by incorporating a separate (logit or probit) equation determining whether or not an individual smokes.

*Duration data.* If the dependent variable is how long it takes to have something occur, such as the number of weeks an unemployed person takes to find a job, a duration model is employed. Estimation is complicated because some of the observations are people who at the time the data were gathered had not yet found work, and so we do not know how long they will take to find work.

## Maximum Likelihood

OLS is a bad estimator for the models listed above (i.e., it has an unattractive sampling distribution). The most popular alternative is the *maximum likelihood estimator* (MLE). The maximum likelihood principle of estimation is based on the idea that the sample of data at hand is more likely to have come from a "real world" characterized by one particular set of parameter values than from a "real world" characterized by any other set of parameter values. The MLE is simply the set of parameter values that gives the greatest probability of obtaining the observed data. This is illustrated in Figure 5.1.

Suppose a random variable $x$ is distributed normally with mean $\mu$ and variance $\sigma^2$. You wish to use the $x$ observations, denoted by small circles on the horizontal axis in Figure 5.1, to estimate the two unknown parameters $\mu$ and

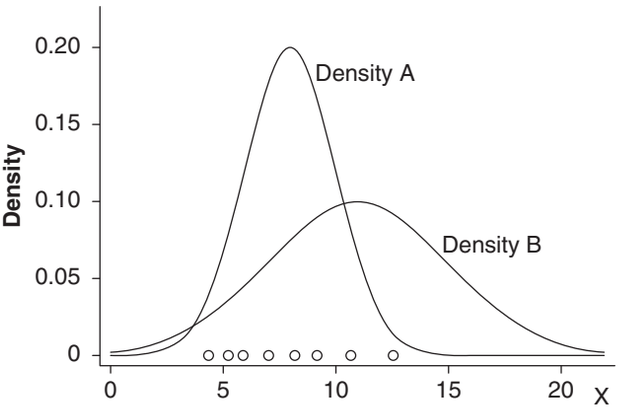


**Figure 5.1** Maximum Likelihood Estimation

$\sigma^2$. The distribution $A$ in Figure 5.1 corresponds to parameter values $\mu_A$ and $\sigma^2_A$. The distribution $B$ corresponds to parameter values $\mu_B$ and $\sigma^2_B$. From looking at Figure 5.1, distribution $A$ is clearly more likely to have produced the observations on $x$, so on the maximum likelihood criterion, we would prefer the parameter values $\mu_A$ and $\sigma^2_A$. By searching over all possible values of $\mu$ and $\sigma^2$, we can find the pair of $\mu$ and $\sigma^2$ values that are most preferred in this sense; these are the MLEs of $\mu$ and $\sigma^2$. In simple cases, an algebraic solution for the MLE can be found; in most cases, a computer search algorithm is used. The MLE has a very attractive sampling distribution, which explains its popularity. In large samples, it is unbiased, it has the smallest variance among all large-sample unbiased estimators, its sampling distribution takes the form of a normal distribution in large samples, and there is a universal formula (called the *Cramer-Rao lower bound*) that can be used to estimate its variance. Its only major drawback is that to calculate the MLE, the econometrician must assume a specific (e.g., normal) distribution for the stochastic component of the model.

## Panel Data

Thanks to the computer revolution, data sets in which we have observations on the same units in several different time periods have become common. These *panel* (or *longitudinal*) data have several attractive features. First, in any cross section, there is a myriad of unmeasured explanatory variables that affect the behavior of the people (firms, countries, etc.) being analyzed. (*Heterogeneity* means that these microunits are all different from one another in fundamental unmeasured ways.) Omitting these variables causes bias in estimation. Panel data enable correction of this problem. Second, panel data create more variability, through combining variation across microunits with variation over time, alleviating multicollinearity problems. With these more informative data, more efficient estimation is possible. And third, panel data can be used to examine

issues that cannot be studied using time-series or cross-sectional data alone. Consider the problem of separating economies of scale from technological change in the analysis of production functions. Cross-sectional data can be used to examine economies of scale by comparing the costs of small and large firms, but because all the data come from one time period, there is no way to estimate the effect of technological change. Things are worse with time-series data on a single firm; we cannot separate the two effects because we cannot tell whether a change in that firm's costs over time is due to technological change or due to a change in the size of the firm.

### Time-Series Data

Some time-series data behave like *random walks*—a variable's value is equal to its previous period's value plus an error term. Such variables are said to have *unit roots.* Because the error terms cumulate over time in this variable, rather than dying out, using such variables in regressions produces misleading results. Differencing the data can remove the unit root and allow regressions to be meaningful but at the cost of throwing away information in the data embodied in their levels—exactly the information economists can provide through characterizing equilibrium relationships. Great progress was made in econometrics by recognizing that this unit root problem could be resolved by combining such variables into equilibrium relationships and estimating a model using differenced data but in which the deviation from equilibrium affects the dynamics of the dependent variable. Such models are called *error correction models;* variables combining into an equilibrium relationship are said to be *cointegrated.*

### Robust Estimation

Econometric theory is very good at devising estimating procedures for special specifications about how the data were generated. But econometricians are seldom in a situation in which they can be confident about their model specification. What if some parts of the specification are wrong? Maybe the errors are more erratic than assumed, for example? *Robust* estimation procedures are designed to be insensitive to violations of the assumptions used to derive an efficient estimator. A simple example is minimizing the sum of absolute errors rather than minimizing the sum of squared errors. *Nonparametric* estimation is a more general form of this, in which estimation proceeds without any functional form constraining estimation.

### Diagnostic Testing

Associated with all topics in econometrics is a need to undertake tests designed to aid specification. Are explanatory variables missing? Is a linear functional form suitable? Are the error variances the same for all observations? Are the errors correlated? Is the error correlated with an explanatory variable? Are there outliers? A big part of studying econometrics is becoming familiar with a wide range of such tests.

## Illustrative Regression Results

Table 5.1 reports a set of OLS regression results representative of what econometric software packages produce. In this example, the logarithm of wage (LNWAGE) has been regressed on years of education (ED) and three dummies, with obvious nomenclature. Because the dependent variable is the log of wage, the estimated slope coefficients can be interpreted as percentages. The computer always prints out results to far more decimal places than is warranted, as illustrated here. The results in the coefficient column indicate that, other things equal, an extra year of education increases wage by about 6%, males earn about 34% more than females, non-Whites earn about 8% less than Whites, and union members earn about 26% more than non-union members. (Because of the nonlinearity here, these numbers are biased; better estimates can be computed as $e^{coeff}-1$. Doing this for the male/female difference, for example, increases the estimate of 34% to about 40%. This adjustment should always be done when the dependent variable is logged and we are estimating the percentage impact of a discrete explanatory variable.)

The *t* statistic column reports the *t* statistic for testing the null hypothesis that its associated coefficient is equal to zero, calculated by dividing the coefficient estimate by its standard error. The Prob. column reports the *p* value for the *t* statistic; this is the probability, if the null hypothesis is true (and assuming a normally distributed error), of obtaining a *t* value bigger in absolute value than the absolute value of the reported *t* value. With the exception of NONWHITE, these probabilities are essentially zero, leading us to reject these null hypotheses. In the case of NONWHITE, this probability is 11%; the decision to reject or not reject rests on what significance level (Type I error) we choose to use for the test. If we choose the popular 5% significance level, we would not reject the null that the coefficient of NONWHITE is zero.

Under the array of coefficient estimates are reported several secondary results of interest, explained briefly below.

- *R-squared* tells us that 28% of the variation in the dependent variable is explained linearly by these explanatory variables. Its adjusted value is only slightly smaller because the sample size (550) relative to the number of explanatory variables (4) is so large.
- *S.E. of regression* is the OLS estimate of the standard error of the error term in this relationship; its square is an estimate of the variance of the error term.

**Table 5.1** Illustrative Regression Results

| *Variable* | *Coefficient* | *Std. Error* | *t-Statistic* | *Prob.* |
|---|---|---|---|---|
| ED | 0.058655 | 0.006643 | 8.829840 | 0.0000 |
| MALE | 0.337946 | 0.037072 | 9.115928 | 0.0000 |
| NONWHITE | –0.077978 | 0.048717 | –1.600616 | 0.1100 |
| UNION | 0.261811 | 0.039135 | 6.689998 | 0.0000 |
| INTERCEPT | 0.668140 | 0.093488 | 7.146809 | 0.0000 |
| R-squared | 0.282894 | Mean dependent var | | 1.681002 |
| Adjusted R-squared | 0.277631 | S.D. dependent var | | 0.490157 |
| S.E. of regression | 0.416596 | Akaike info criterion | | 1.095649 |
| Sum squared resid | 94.58586 | Schwarz criterion | | 1.134830 |
| Log likelihood | –296.3034 | Durbin-Watson stat | | 1.955034 |
| F-statistic | 53.74976 | Prob(F-statistic) | | 0.000000 |

NOTES*:* Dependent Variable: LNWAGE. Method: Least Squares. Sample Size 550.

- *Sum squared resid* is the minimized value of the sum of squared residuals resulting from the OLS procedure.
- *Log likelihood.* If the error term is assumed to be distributed normally, the MLE, obtained by maximizing the likelihood of the sample (the probability of obtaining the sample), is the OLS estimator. In practice, the MLE is found by maximizing the log of the likelihood. *Log likelihood* is this maximized value.
- *F-statistic* reports the *F* statistic for testing the null hypothesis that the slope coefficients are jointly all equal to zero. Its numerator degrees of freedom is the number of individual hypotheses being jointly tested, in this case 4; its denominator degrees of freedom is the sample size less the number of coefficients being estimated, in this case 545.
- *Prob(F-statistic)* is the *p* value of this *F* statistic, the probability, if the null is true, of obtaining an *F* value bigger than the reported *F* value.
- *Durbin-Watson stat* is a test statistic used for testing for autocorrelated errors when using time-series data. If it is close to 2, the null of no autocorrelation is not rejected. Because the wage data are cross-sectional, this number is without meaning in this example.
- *Akaike info criterion* and *Schwarz criterion.* It is tempting to select the set of explanatory variables by maximizing adjusted $R^2$, implying that in essence one chooses the specification that minimizes the sum of squared errors plus a penalty for using extra regressors. The Akaike information criterion (AIC) and the Schwarz criterion (SC) are competing measures that employ "better" penalties for using extra regressors.
- *Mean dependent var* is the average of the observations on the dependent variable; *S.D. dependent var* is the standard deviation of these observations.

## References and Further Readings


Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*. Norwell, MA: Kluwer.

Belsley, D. A. (1986). Model selection in regression analysis, regression diagnostics and prior knowledge. *International Journal of Forecasting, 2,* 41–46, and commentary pp. 46–52.

Berndt, E. R. (1991). *The practice of econometrics: Classic and contemporary.* Reading, MA: Addison-Wesley.

Darnell, A. C. (1994). *A dictionary of econometrics.* Aldershot, UK: Edward Elgar.

Dewald, W., Thursby, J., & Anderson, R. (1986). Replication in empirical economics. *American Economic Review, 76,* 587–603.

DiNardo, J., & Tobias, J. L. (2001). Nonparametric density and regression estimation. *Journal of Economic Perspectives, 15,* 11–28.

Engle, R. F. (2001). GARCH 101: The use of ARCH/GARCH models in applied econometrics. *Journal of Economic Perspectives, 15,* 157–168.

Geweke, J. K., Horowitz, J. L., & Pesaran, M. H. (2008). Econometrics: A bird's eye view. In L. Blume & S. Durlauf (Eds.), *The new Palgrave dictionary of economics* (2nd ed.). London: Macmillan.

Griliches, Z. (1985). Data and econometricians: The uneasy alliance. *American Economic Review Papers and Proceedings, 74,* 196–200.

Kennedy, P. E. (2002). Sinning in the basement: What are the rules? The ten commandments of applied econometrics. *Journal of Economic Surveys, 16,* 569–589, with commentary and reply.

Kennedy, P. E. (2005). Oh no! I got the wrong sign! What should I do? *Journal of Economic Education, 36,* 77–92.

Kennedy, P. E. (2008). *A guide to econometrics* (6th ed.). Malden, MA: Blackwell.

Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics, 65,* 1–12.

Magnus, J. R. (1999). The success of econometrics. *De Economist, 147,* 55–71.

Masten, S. E. (2002). Modern evidence on the firm. *American Economic Review Papers and Proceedings, 92,* 428–432.

Mayer, T. (1980). Economics as a hard science: Realistic goal or wishful thinking? *Economic Inquiry, 18,* 165–178.

McCullough, B. D., & Vinod, H. D. (1999). The numerical reliability of econometric software. *Journal of Economic Literature, 37,* 633–665.

Murray, M. P. (1994). A drunk and her dog: An illustration of cointegration and error correction. *American Statistician, 48,* 37–39.

Murray, M. P. (2006). The bad, the weak, and the ugly: Avoiding the pitfalls of instrumental variable estimation. *Journal of Economic Perspectives, 20*(4), 111–132.

Pagan, A. R. (1987). Three econometric methodologies: A critical appraisal. *Journal of Economic Surveys, 1,* 3–24.

Smith, R. P. (1998). Quantitative methods in peace research. *Journal of Peace Research, 35,* 419–427.

Swann, G. M. P. (2006). *Putting econometrics in its place: A new direction in applied economics.* Cheltenham, UK: Edward Elgar.

Ziliak, S. T., & McCloskey, D. N. (2004). Size matters: The standard error of regressions in the *American Economic Review. Journal of Socio-Economics, 33,* 527–546.

# 6

# Marxian and Institutional Industrial Relations in the United States

Michael Hillard

*University of Southern Maine*

This chapter presents industrial relations (IR)—the study of the capitalist employment relationship, with particular emphasis on employer/worker or "capital-labor" conflict. The field originated a century ago in the United States, maintaining an intellectual tradition and historical experience distinct from IR in other nations. The field was almost a purely American tradition until after World War II, when it was promoted and exported to Europe and developing countries, with U.S. cold war foreign policy an important impetus. The intellectual landscape and character of IR in other nations, with Britain a singular example with its heavy Marxist influence, is completely distinct and beyond the scope of this chapter. Kaufman (2004) provides a useful global history of IR.

Two IR schools of thought are summarized—the original, *institutional* tradition of industrial relations (or ILE/IR for institutional labor economics/industrial relations) and *Marxian-derived* modern political economy, which, drawing on the work of historians of U.S. labor and basic Marxian thought about class relations, has a competing interpretation of both the employment relationship and class conflict. Both schools provide an interpretation of U.S. labor relations' history and a policy agenda for solutions to employment relationship problems.

It is important to note that institutional and Marxian analyses share important conceptual ground. They are both *heterodox* theories (i.e., they are outside of the post-1950s Anglo-American neoclassical mainstream). Both emphasize the social determination of human beings' consciousness and norms of economic behavior (i.e., as opposed to seeing people as individual maximizing agents with preferences given and prior to any social influences), with institutions and power fundamentally shaping individual behavior and group economic outcomes. Most important, they share a belief that the default "market" situation in a "free" labor market is one of unequal bargaining power of capital over labor, an inequality that provokes conflict. They differ on understanding the nature and sources of power, especially in their interpretations of class conflict and of appropriate policy. This will all be explored in detail below.

Finally, this chapter presents IR's historical background. Class conflict is a historical phenomenon, with ebbs and flows and patterns of evolution and devolution. The origins of the field, the late coming of radical economic theory to the stage, and even a grasp of contemporary employment relations questions, in the view of the field, can be understood only historically. So, to situate our comparative analysis of institutional and Marxian schools of IR theory, we begin with the origins of IR in the industrial conflict that grew out of the American Industrial Revolution. Comparisons of the ILE/IR and Marxian views of the employment relationship, IR policy, understanding the decline of labor in the late twentieth century, and the future direction of the field follow this section.

## Capitalism and the Labor Question in the United States

It is widely understood by labor experts that industrial revolutions—the rise of a specifically capitalist system based on wage labor—create a brew of social and economic

conditions that provoke labor conflict (see Cowie, 1999; Dunlop, Kerr, Harbison, & Myers, 1960; Marx, 1977). The U.S. Industrial Revolution began in textiles in the early nineteenth century and gained full force after the Civil War when a new national railroad system sparked the rapid growth of a national market.

By the 1880s, the strongest force in the lives of the quickly growing U.S. industrial working class was their employer. Workers employed by large and small industrial employers who faced fierce competition in a national market—the world's largest—experienced harsh conditions, including low wages, long hours, management authoritarianism, and a lack of civil liberties in industrial communities. Early institutional thinkers such as Richard Ely and John R. Commons began to label these collectively as *labor problems* (Kaufman, 1993). As in all industrializing societies, American workers began rebelling, forming oppositional movements with both *militancy* (strikes, organized restriction of output, etc.) and *radicalism* (various ideologies of revolution within the labor movement) that were a threat to both the ability of capitalists to earn a profit and a potential long-run threat to capitalism itself. In turn, extreme hours, workplace exploitation, and chronic unemployment afflicted and provoked workers to organize and rebel.

As early as the 1870s, American workers created unions, joining in large numbers, in order to improve their chances of reducing hours (specifically the 8-hour day), gain fairness in the workplace and in their communities, and increase wages above meager levels. National railroad strikes in 1877 and 1894; nationwide marches and protests by workers on May 1, 1886, for the 8-hour day; and the famed Homestead steel strike in 1892 all seized the nation's attention. All were brutally repressed by state and federal troops. By the 1910s, the United States experienced an unprecedented, violence-filled strike wave that spread beyond "native" workers to include southern and eastern European immigrants, the growth of socialist politics and anarchist unionism, and acts of violence and terrorism that were seen as threatening prosperity and even social stability.

These phenomena, and the desire to bring about a progressive resolution to them, came to be known as *the labor question.* As President Woodrow Wilson framed it in 1919, the labor question was as follows:

> How are the men and women who do the daily labor of the world to obtain progressive improvement in the conditions of their labor, to be made happier, and to be better served by the communities and industries which their labor sustains and advances? (quoted in Lichtenstein, 2002, p. 4)

The labor question was highly political because of the degree of social conflict and strongly opposing ideologies associated with the different sides. The first choice of most employers was to use the government as an instrument of repression, along with spies and private militias. For instance, state militias were brought in to suppress strikes at least 495 times between 1880 and 1900, and by the 1930s, more than 100,000 were employed in union-busting "detective" agencies (Green, 1998; Laurie, 1989). This lasted until a sea change in government involvement and societal attitudes attenuated widespread use of overt repression of labor militancy roughly in 1940.

Political leaders, journalists, labor activists, and some business leaders responded to accelerating labor conflict in the 1910s by searching for labor problem solutions that addressed root causes and did not further inflame labor strife by breaking unions. Among experts, public leaders, and the union movement itself, the term *industrial democracy* became the term that described these alternatives to business repression and overt class warfare. Three visions of industrial democracy emerged during this era, becoming the contending models for IR policy for much of the twentieth century.

New employment relations experts defined two of these views—a "personnel management" (PM) tradition and the ILE/IR tradition (Jacoby, 1997; Kaufman, 1993; Kochan, Katz, & McKersie, 1986). These were clear and contending *professionally* driven visions and practices of how to bring about labor peace—an anti-union, PM tradition and the pro-union (though antiradical unionism) ILE/IR tradition. IR experts sought to enlist the state to support moderate "business unionism" (defined below) and create a regulatory role for IR experts themselves, while the PM group saw itself as a component of management. Personnel management was continually in direct competition with the authoritarian approach favored by most employers, while ILE/IR sought national government policies.

Within the labor movement itself, there was a split between proponents of business unionism and radicals. The American Federation of Labor (AFL) was the dominant national union institution. Founded in 1886, it was a federation composed of existing craft-based trade unions. The AFL's predominance was unchallenged until the formation of the Congress of Industrial Organizations (CIO) in the mid-1930s. The AFL advocated *business unionism,* which sought to have employees collectively bargain directly with employers for improvements (via compromises embodied in contracts) in wages and working conditions. The AFL did not seek a major role for government in industrial relations other than to stop employers and courts from repressing their rights to associate and bargain. The AFL position was consistent with and connected to the IR school's vision of employment relations. A third vision of industrial democracy came from a growing radical wing of the 1910s' labor movement that embraced both socialist politics and a more militant unionism favoring strikes and direct action.

While use of the phrase *industrial democracy* eroded in the coming decades, the basic categories of employment relations' policies continued throughout the rest of the

century, with some more ascendant in certain periods than others. Radical unionism/socialist politics persisted until its full repression in the 1940s, and the "non"-solution of employer autocracy remained the default practice for a majority of employers.

The U.S. federal government, needing employment stability during World War I, briefly instituted a version of the IR solution during 1917 and 1918. It was quickly dismantled after the war, and labor unions, which had grown rapidly, were just as quickly crushed. During the 1920s, large- and medium-sized U.S. employers hired PM experts widely to enact "corporate welfarism." Employers with as many as 3 million employees came under the aegis of "welfare" policies that included new disability and pension benefits, improved management practices, and "company" (i.e., employer created and dominated) unions (Brody, 1993; Jacoby, 1997).

The Great Depression brought about the ascendancy of the ILE/IR expert group, as the policies offered via the PM model collapsed in the economic crisis, leaving only employer autocracy, which worsened the depression by reducing wages and buying power. The IR model was embraced and sponsored by key congressional Democrats and the Roosevelt administration. They were its designers, rank-and-file bureaucrats, scholarly analysts, and public relations champions. ILE/IR members were both academics and political reformers devoted to a centrist solution to the problem of class conflict. In the 1930s, Democrats embraced ILE/IR proposals as a major component of liberal economic policy. IR policy came to the fore because of the peculiar and deep crisis of the 1930s (Kaufman, 1993; Lichtenstein, 2002). Crucially, IR policy offered a solution to the depression itself because it would lead to rising industrial wages and thus support greater mass consumption (i.e., effective aggregate demand) that would in turn revive the economy. It thus was a pillar of what came to be Keynesian macroeconomic policy. Also, IR policy offered an alternative to the twin "evils" of socialism and class warfare.

Two signal and roughly coterminous events marked the mid-1930s: the passage of the National Labor Relations Act (NLRA) in 1935 and the rise of the CIO, a federation of unions composed of populist and left-wing *industrial* unions. *Industrial* connotes a union composed of all workers in a workplace without regard to occupation or "trade." The AFL is a federation composed of *trade unions*—that is, with membership based principally on occupation (e.g., carpenters, machinists, typographers). The AFL bitterly fought industrial unionism from its inception through the 1930s (Dubofsky, 1996). CIO unions organized for the first time the preponderance of workers in twentieth-century mass production industries (steel, auto, rubber, electrical workers, etc.). The NLRA, designed by ILE/IR experts, legalized unions and created a government-sponsored election process to allow workers to elect for union representation, a requirement that capital and labor bargain collectively "in good faith," and, most important, proscription of traditional employer "unfair labor practices" (e.g., firing and blacklisting pro-union employees) that had previously blocked widespread unionization. New organization and the support of the federal government, combined with heavy pressure by government on recalcitrant industrial employers during World War II, sparked a fivefold expansion of union membership from 1933 to 1945. This ushered in the "New Deal IR system." After thriving for decades, the New Deal system went into an abrupt decline in the 1980s. The New Deal system provided dramatic increases in working-class living standards (however, with many groups left out—southern and some rural workers, African Americans, and women workers). Its end in the period around 1980 reversed this progress.

## The Institutional IR School: Peculiarities of the Labor Market, Unequal Bargaining Power, and Capital–Labor Conflict

IR began as a subset of the institutional economics field, focused on the question of capital–labor conflict and compromise solutions to that conflict. Institutional economists developing IR theory also studied the need for social insurance and sought to understand the institutional forces shaping the labor market.

Beginning with Richard Ely and especially with the work of John R. Commons and his students in the Wisconsin school in the late nineteenth and early twentieth centuries, an evolutionary line of descent can be traced through a second generation of scholar/policy makers from the 1930s to the 1990s, led by John Dunlop and Clark Kerr, Richard Lester, and Arthur Ross. This group was somewhat friendlier to orthodox (neoclassical) economics. A third generation is active and influential today (including Kate Bronfenbrenner, Michael Piore, Richard Freeman, Eileen Applebaum, Thomas Kochan, and Bruce Kaufman). Some in this third generation differed from the second in their greater concern for problems of urban poverty, racism, and sexism.

Richard Ely began with a view of labor markets that differed radically from the assumption of "free" labor markets found in neoclassical economics. Ely concluded that capitalist labor markets suffered three "peculiarities" that disadvantage workers vis-à-vis employers: a lack of bargaining power, management's authoritarian treatment of workers, and economic insecurity—that is, frequent unemployment (Kaufman, 1993, pp. 32–33). If labor in capitalism is "free" (i.e., free to quit a bad employer and find another, creating a competition among employers that should redress such problems), how could employers have such an upper hand? The premise is significant, chronic unemployment. Beatrice Webb (1901) provides a clear statement of the sources of this imbalance that defines why

capitalist labor markets were then and now are slanted in employers' favor:

> If the capitalist refuses to accept the workman's terms, he will, no doubt, suffer some inconvenience as an employer. . . . But, meanwhile, he goes on eating and drinking, his wife and family go on living, just as before. His physical comfort is not affected: he can afford to wait until the labourer comes back in a humble frame of mind. And this is just what the labourer must presently do. For he, meanwhile, has lost his day. His very subsistence depends on his promptly coming to an agreement. If he stands out, he has no money to meet his weekly rent, or to buy food for his family. If he is obstinate, consumption of his little hoard, or the pawning of his furniture, may put off the catastrophe; but sooner or later slow starvation forces him to come to terms. And since the success in the haggling of the market is largely determined by the relative eagerness of the parties to come to terms—especially if this eagerness cannot be hidden—it is now agreed, even on this ground alone, "that manual labourers as a class are at a disadvantage in bargaining." (pp. 8–9)

In turn, these resulted in labor *problems.* In distinction to the "labor question," *labor problems* included workers' low wages, insecure employment/income, and harsh treatment by supervisors. Employers faced the wake of workers' self-protective reactions, ranging from lost output due to workers' restriction of output and lost productivity from strikes. Society suffered because of lost production but mostly the disruption of violent strikes and nationwide strike waves (Kaufman, 1993).

Finally, conflict was inevitable in the face of such harsh conditions. The militancy and radicalism of industrial nations' workers seeking to redress exploitation and insecurity was well established by the late 1800s. As discussed below, this led to a policy vision focused on using the state and harnessing an invigorated, state-supported, conservative trade unionism to bring about labor peace through a government-supported process of compromise.

## Marxian IR: Marx and Capitalist Employment Relationship

Marx developed a theory of the inherent presence of class conflict in capitalist employment relations. Radical and Marxian analyses of the U.S. thrived until the early cold war period, when McCarthyist repression of radical intellectuals reduced radicalism to a marginal existence. The changed political environment of the 1960s permitted a revival. Scholars such as Samuel Bowles, James Crotty, Heidi Hartman, and Stephen Resnick and Richard Wolff led a rebirth around 1970. Marxian labor analysis grew out of the influential work of Harry Braverman (1974) and Gordon, Edwards, and Reich (1982).

Marx begins with what he sees as the historic fact of classes. Capitalism's class system is seen as the product of two related historic processes: "proletarianization" and primitive accumulation (Marx, 1977, pp. 873–930). Proletarians, capitalism's working class, are formed out of former serfs, independent "yeoman," or peasant agricultural workers who were forcibly separated from their access to land and dispossessed in European societies. Paupers moved to cities and sought waged employment. These proletarians were newly "free" to sell their labor time (for Marx, a person's "labor-power") and were "freed from" access to the land and resources needed to produce their means to survive. They are thus both "free" to participate in labor markets and, at the same time, utterly dependent on them. In turn, European aristocratic classes participated in global plunder and exploitation of slave-produced commodities such as sugar, as well as the slave trade, to produce a merchant class with the money to hire proletarians and begin the process of manufacturing for markets for a profit (Howe, 2002). Capitalists have exclusive control and access to physical capital (which Marx termed the "means of production") and money that can provide workers with a means of buying goods and services—thus being able to survive. Workers are compelled under these circumstances to work for a wage and accept giving over their time to the control of capitalists. The drive for profit by capital, as well as the control they have over workers' lives, results in exploitation and, in the absence of worker resistance or government limitations, tremendous excesses such as child labor and extremes of workdays (pushing physical limits of 14 to 18 hours per day) and working conditions that inevitably shortened life spans.

Marx published Volume I of *Capital* 20 years prior to Ely's founding of U.S. institutional economics, a history consistently ignored by IR scholars in the United States. Yet, Marx's analysis identified Ely's "three peculiarities of the labor market" and went beyond Ely to highlight workers' exploitation.

For Marx, the wage is set just like any other value, at the cost of the reproduction of the thing being sold. Thus, in Chapter 6 of *Capital,* Volume 1, Marx presumes that the labor market is fair in the precise sense that the wage is equal to what it costs to clothe, feed, house, and transport the worker and provide for the next generation at a socially and customarily determined standard of living. It is not in the labor market but *inside* the employment relationship that one must go to understand processes of conflict and cooperation.

What the capitalist wants to buy is labor (i.e., work), but what he or she can buy is labor time. For radical economists, this is Marx's famous "labor from labor-power" distinction. Capitalists buy from workers a commodity: the human ability to apply muscle and brain power. The cost of this commodity is set by its labor time necessary to produce (Marx used a "labor time" theory of commodities' value or price). Marx argued that a given capitalist society would have a socially determined

(i.e., by norms established historically and often through class struggle) combination of "wage goods." The labor time required to produce this "wage bundle" (i.e., what is necessary to socially reproduce the worker and the next generation of workers) would define the wage as the equivalent of "necessary labor time" (i.e., keeping a workforce alive and able to continue economic production). Capitalists earn a profit by getting more labor time from a worker than the wage. Labor, unlike the fixed cost of labor-power, is elastic and can be extended or intensified to increase labor time embodied in newly produced commodities beyond the cost of the wage, and this constitutes "surplus labor time." Surplus labor time is the source of profit (i.e., surplus value).

The labor from labor-power distinction causes two kinds of conflict. First is conflict over the length of the working day. If capitalists can lengthen the working day without increasing the (daily) wage, more of the output returns to them: that is, surplus labor time and the surplus product (i.e., the labor time and production above "necessary labor time"). Competition drives capitalists to push workers, in the extreme, to work arduous hours—up to and past 80 hours a week, to the point where workers' health is damaged and lives are shortened. Indeed, the beginning of the modern labor movement can be found in the struggle to limit the working day, first for women and children and then for all workers.

A second area of conflict is reduction in necessary labor time, or what Marx called *relative surplus value.* Here, the object is to reduce necessary labor time through increased productivity. Individual capitalists have an incentive to pursue this, in order to increase profits at the expense of competitors through cost advantages. Technical changes that increase the pace of production or allow skilled or high-status workers to be replaced by unskilled or low-status workers, production speedup that gets the worker to work harder, and moving production to regions with a lower standard of living all increase the relative part of the product that accrues to the capitalist. At the center of this strategy is intensification of the work pace beyond human limits and de-skilling of work that destroys the intrinsic value of skilled work. Both provoke worker resistance and rebellion.

While workers may struggle (sometimes successfully) against these two forms of exploitation, their success is limited by a number of factors, including the problem of coordinating individual action for mutual gain and the existence of the reserve armies of the unemployed who are ready to take employment under lower standards than incumbent workers.

Marx illustrated his ideas in a detailed case study of Britain's Industrial Revolution, with its struggle over the length of the working day, the rise of mass production, and the intense competition that makes capitalism both technically progressive but also intensely exploitative. British capitalists, seeking an advantage in competition, are shown pushing the working day past all physical and moral limits, including cruelties that dramatically shorten the lives of workers and rob children of a meaningful childhood. Marx (1977) argues that the incentive for every capitalist to use up labor-power without worrying about there being a tragedy of the commons can be limited only by society itself placing absolute constraints on the capitalist:

> *Apres moi le deluge!* is the watchword of every capitalist and of every capitalist nation. Capital therefore takes no account of the health and the length of life of the worker, unless society forces him to do so. (p. 381)

Workers, seeking to preserve their "property" (themselves) and their humanity, demand a working day that does not physically destroy them and leaves time for cultivation of their physical, intellectual, and cultural powers. Similarly, an inherent conflict exists over the *intensity* of labor, particularly once society establishes some norms on the extent of the working day. Capitalists divide the labor process, create single-task jobs (the "detail worker"), and generally impoverish the work process, increase its speed, and, by using specialized machinery (in what subsequently came to be known as "mass production"), create a super-productive work process that at the same time degrades work and the worker. As the "new" labor history and labor process theory has shown in recent decades, U.S. workers have indeed periodically and frequently rebelled over the length of the working day and over "control" in the labor process (see, e.g., Braverman, 1974; Montgomery, 1980; for a careful review of 30 years of "labor process" work provoked by Braverman's book, see Thompson & Newsome, 2004).

The institutional and Marxist traditions thus share the common assumption that class conflict under capitalism is to be expected. Where they depart ways is on what is the appropriate *resolution* of this conflict. For institutionalists, class conflict is seen as a threat to social order, a "primitive democracy" to be channeled and ultimately repressed (S. Webb & Webb, 1897). Marx proposed a different notion.

Marx infamously predicted in the *Communist Manifesto* (coauthored with F. Engels) a worker-led revolution that never materialized (Tucker, 1978, pp. 473–500). Marx also argued that the working class could, through waging class conflict, improve its position within an existing economic system *without* overthrowing it. He illustrated this idea by describing the British working class's victorious struggle to shorten the working day and reduce child labor:

> It will be easily understood that after the factory magnates had resigned themselves and submitted to the inevitable, capital's power of resistance gradually weakened, while at the same time the working class's power of attack grew with the number of its allies in those social layers not directly interested in the question. Hence the comparatively rapid progress since 1860. (Marx, 1977, pp. 408–409)

Marx (and Engels) stressed that even within the existing structure of capitalism, the state *could* be pressed by circumstance into opposition to the interests of capitalists. Friedrich Engels, Marx's collaborator, makes this point in his famous *Letter to Bloch,* citing the many occasions in which they made this point:

> If Barth therefore supposes that we deny any and every reaction of the political, etc., reflexes of the economic movement upon the movement itself, he is simply tilting at windmills. He has only got to look at Marx's *Eighteenth Brumaire,* which deals almost exclusively with the *particular* part played by political struggles and events; of course, within their *general* dependence upon economic conditions. Or *Capital,* the section on the working day, for instance, where legislation, which is surely a political act, has such a trenchant effect. (Tucker, 1978, p. 765)

A politically empowered working-class labor movement thus could and has shaped capitalist societies to improve treatment of the working class. Marx's conclusion about the struggle over the length of the working day is instructive on "reformist" possibilities within capitalism:

> For "protection" against the serpent of their agonies, the workers have to put their heads together and, *as a class,* compel the passing of a law, an all-powerful social barrier by which they can be prevented from selling themselves and their families into slavery and death by voluntary contract with capital. (Marx, 1977, p. 416)

The moments when the state leans against the interests—espoused and practical—of the capitalist class are restricted to those conjunctures when a combination of "pressure from below" and a host of other circumstances—political, economic, and ideological—combine to produce a change in the state that reacts back on the economy. Marx, despite his infamy as a predictor of the necessity and inevitability of the revolutionary demise of capitalism provoked by a combination of impoverishment of all workers and irresolvable economic crisis, had, with Engels, a specific vision of reform within capitalism that could better the fortunes of the working class.

## ILE/IR: A Program for Labor Peace and an End to Labor Problems

ILE/IR scholars defined a clear paradox: how to equalize the bargaining power of labor (thus curbing the "peculiarities of the labor market" and especially creating equal bargaining power between employer and worker) and *not,* as a consequence, provoke an empowered working class into more militancy and creating even more capital–labor conflict. ILE/IR offered a reform vision that, if implemented successfully, would resolve this paradox, bringing stability and steady economic growth. It called for "corporatism," a system where employer and worker representatives work together through a collective bargaining process to produce capital–labor stability by means of compromise. Just as important, the ILE/IR school promoted the nonrevolutionary business unionism of the AFL as the preferred mode of making corporatism work. The AFL and IR policy together provide *an alternative to radicalism,* defeating radicalism by making it irrelevant.

Kaufman (1993) summarizes this ethos in describing the ILE/IR corporatist view of "the efficacy of conflict in the employment relationship":

> From the point of view of the institutionalists, a certain amount of conflict is the normal by-product of the employment relationship and, indeed, frequently plays a constructive role to the extent that it vents repressed frustrations, resentments, and grievances. Good industrial relations, therefore, is not synonymous with an absence of conflict, for often this indicates complete domination of the relationship by the employer. Rather, good industrial relations requires equalizing the bargaining power of labor and capital both inside and outside the plant and letting them voluntarily negotiate a mutually satisfactory outcome. The watchword of the institutionalists is compromise. (p. 38)

Such compromise is imagined to be at least partly dependent on the work of state-supported mediation experts. Indeed, it is important to note how far this vision differs from a free-market view of the economy. Institutions—including unions and employers with bureaucratic representatives and working within parameters set, enforced, and often directly supported by the state—structure the process of striking a bargain, rather than "free," individual agents in a market.

More broadly, about the purpose and result of creating such a state machinery of IR experts, Commons famously said, "In dealing with the momentous conflict of 'capital and labor' . . . I was trying . . . to save Wisconsin and the nation from politics, socialism, or anarchy" (cited in Ramirez, 1978, p. 188). In other words, IR experts were to provide not just workplace stability but also political stability—making labor revolt and anticapitalist working-class political expressions less likely. In its place, Commons advocated for a system in which "neutral" experts provided technically determined compromises between the conflicting interests of capital and labor.

IR scholars have thus had as their goal the careful management of class struggle, with a fundamental aim of limiting (though not suppressing) conflict and encouraging nonsocialist political expression by workers through nonpartisan, nonmilitant unions. In turn, their role as experts has been to continuously study the underlying "problems" that might cause an unstable landscape—notably through case studies of industrial relations in specific industries—and get directly involved as mediators and experts, to ensure that the problems will not get out of hand and that compromises are struck. IR experts saw, for the New Deal

IR period, what Kaufman terms *institutionalization* in the creation of IR institutes with the creation of graduate programs at the top state universities throughout all regions in the United States except the South and also in federal and state mediation bodies.

It appeared that ILE/IR policy worked in the 1950s and 1960s. The mass strike activity of the 1930s/1940s abated. The U.S. economy experienced a combination of unparalleled income "compression" (i.e., greater equality) that saw industrial workers become middle class and enjoy steady and powerful income growth, as well as improvements, including fewer hours worked and an expanded private sector safety net provided through union contracts. The ILE/IR field itself appraised this as the success of expertise over a technical problem with capitalism and projected that the stability of the New Deal system would persist indefinitely. However, through the exercise of capitalist class power, this system came to an abrupt end in the 1980s.

## Marxian Political Economy: Class Strength and Workers' Improvement *Without* Revolution

Marxian analysis of U.S. employment relations employs a different analytical framework. It comes to several different conclusions that contrast and even contradict the ILR/IR school:

- Not only is class conflict inevitable, but it can be a good thing.
- A stronger labor movement tends to produce better social outcomes, including lower income inequality, less poverty, shorter weekly and annual work hours, and more comprehensive social insurance, without an overthrow of the capitalist system.
- Historical evidence supports these conclusions.

It has been the political quest of working-class parties in advanced capitalist countries to accomplish that very goal of improving the quality of life for workers and for society, seeking to humanize capitalism by minimizing economic insecurity and discrimination, reducing/eliminating poverty, implementing safety and health and labor standards (e.g., limiting hours worked, eliminating child labor), and providing public goods (e.g., education, child care, public space). The labor movements of various countries have also sought regulatory and union agreements that grant expanded workers' rights and decreased property rights, such as legal guarantees over right to job, workplace decision-making power, and having a significant stake in corporate governance.

Thus, Marx's emphasis on relative class strength as a predictor of social outcomes, as well as the possibility of a politically strong working class to enlist social allies (e.g., middle classes) and the state, has resonance with twentieth-century historical experience. Succinctly, the quality of economic and social life for a working class in a particular nation is a reflection of the strength and degree of past and current success of its labor movement. Child care, retirement, health care, social security, length of the working week, and weeks of paid vacations are better for workers in Sweden, Germany, or France than in the United States. Why? Because over the second half of the twentieth century, their labor movements succeeded economically and politically where the U.S. labor movement either failed or won only modest improvements (and, as noted, isolated only to certain regions) that a now far weaker labor movement has been unable to protect. There is of course no reason to believe that these improvements will lead to revolutionary change, but such improvements are, we would argue, good in and of themselves. Added to Marx's historical example of the 10-hour day, this offers compelling evidence that a politically and economically stronger working class, as well as class conflict emanating from it, produces positive outcomes for the working class. When two valid and opposite claims over the course of the employment relationship meet one another, as Marx put it, in the end, force decides.

To reiterate, a politically strong working class can, with allies and supportive historical circumstances, get the state to "lean against" the interests of the capitalist class. This happened in the United States in 1934–1937.

## The Case of 1934–1947

The Great Depression reignited working-class insurgency after more than a decade of dormancy. By 1934, workers, emboldened by Roosevelt's verbal support for unions, mobilized to form unions and conducted vigorous strikes, including three "general" (community-wide, multiemployer) strikes (Green, 1998; Lichtenstein, 2002; Lynd, 1996). Initially, Roosevelt's National Recovery Act implemented in 1933 and 1934 large businesses' proposal for solving the depression. This consisted of government-sponsored cartelization—businesses openly cooperating in setting prices and allocating markets. This approach did not address the underlying problem of inadequate aggregate spending and utterly failed. In combination with the public's already low esteem for big business—the American public held business responsible for causing the Great Depression in the first place—this failure left a political vacuum seized upon by two allied groups: the broad working-class insurgency played out in streets and factories across the country, and liberal political leaders who sought to impose ILE/IR policies such as the NLRA and a broad program of social security. This magnified the already existing loss of business legitimacy that the Great Depression created, opening the way for a coalition of congressional and executive branch leaders, representing the northern urban, working-class, union-based movement, to seize control of national policy making.

This permitted a onetime imposition of values and policies alien to U.S. capitalists. The loss of capital's credibility and power was unprecedented and short-lived, but it ushered in dramatic changes—an expanded welfare state and widespread unionization—that lived on for decades (Finegold & Skocpol, 1984; Lichtenstein, 2002). The results: a New Deal state that imposed state-sponsored support for an IR collective bargaining–based system (under the 1935 Wagner/NLRA Act), and a tax-supported national social welfare system, including Social Security and unemployment insurance, along with significant labor market regulation (especially the 1938 Fair Labor Standards Act). During this period, industrial unionism caught fire, gaining special momentum after Roosevelt's 1936 reelection; the new political environment contributed in part to high-profile union victories at General Motors and U.S. Steel in early 1937. While interrupted briefly by the late 1930s recession, the momentum toward rapid unionization resumed during World War II and continued into the 1950s.

In sum, this period illustrates a similar moment of dramatic, pro–working-class reform, supported and implemented to a significant degree by a state willing to act against capital's economic and political interests. Working-class political strength was unparalleled and briefly even unchallenged by capital. The result was long-lasting improvements for the working class.

This observation comes with an important qualification. Historical scholarship has a consensus view that African Americans, southern workers in general, and women were largely left out of this progress. At the level of policy (e.g., Social Security and unemployment coverage, as well as wages and benefits), this was most certainly true. Many in these groups were trapped in a "secondary" labor market of low wages and benefits with little opportunity for income security or advancement (Gordon et al., 1982). There are also notable exceptions to this partial progress (Minchin, 2001). It is also the case that new activism in the 1960s and 1970s served to expand inclusion of some women and people of color in some of these benefits, but even then incompletely (Kessler-Harris, 2007; Lichtenstein, 2002).

## ILE/IR: Stagnating New Deal System Collapses, Dynamic Non-Union Alternatives Rise

Recently, ILE/IR scholars have reinterpreted twentieth-century U.S. IR history, focusing on explaining why the dominant unionized/corporatist New Deal system lasting from the 1930s through the early 1970s went into an abrupt and lethal decline (Jacoby, 1997; Kaufman, 1993; Kochan et al., 1986). They attribute the decline to the vitality of employer-dominated "welfare capitalism" throughout the era of the New Deal, which escaped the attention of earlier IR scholars. The unexpected vulnerability of unionized employers to rapid economic change from the 1970s led to a new dominant IR system based on "progressive" but union-free employers.

How could a strong and durable institutional system be taken apart at all, much less so quickly? First, 1920s welfare capitalism survived (and adapted) in the 1940s–1970s period, becoming more sophisticated, and gained strength and momentum. Non-union employers also led capitalist political activism against the New Deal state and unions. For instance, the 1948 Taft-Hartley Act created vast loopholes in labor law that allowed employers to defeat new union organizing drives. In retrospect, the true apogee of union strength was really in the 1950s. From then on, dynamism in IR policy centered on the ever-expanding tool set of welfare capitalism—the latest version of the "PM tradition" now called human resource management (HRM) that contributed to the vitality of non-union companies. Even unionized companies adopted HRM practices in efforts to weaken union influence (Phillips-Fein, 2009). Besides high wages and generous benefit packages—aimed to meet the standard set by the strongest union contracts—the HRM movement sought to realize the PM vision of a humane management by surveying workers to gain a grasp of their concerns, training frontline supervisors to be more positive, and creating at least an appearance of due process through employee handbooks and "open-door" policies. As we discuss below, the velvet glove of HRM shielded an iron fist that crushed any movements toward independence through union organizing. Union busting, capital flight to non-union states or offshore locals, and "progressive management" were part of one package.

When rapid globalization met the macro crisis of 1979–1983, unionized business abandoned its cooperative relationship with unions, getting givebacks, deindustrializing much of the Midwest, and, with perhaps the singular exception of auto, shifting more aggressively to non-union operations in Southern and Great Plains rural towns or "offshore" to developing countries with low wages and labor standards. Another factor was simple contraction in the face of loss of markets to foreign competition, where companies ceded markets rather than investing in new technology and capital, instead diverting retained earnings into conglomerate purchases, as was the case in steel. Private sector union density (the unionized percentage of workers) fell by two thirds over the last three decades of the twentieth century, standing now below 8%.

## A Marxian View: Changing the Terms of Class Power, American Employer Exceptionalism

To reiterate, the central concept in a Marxian interpretation of industrial relations is relative class power. Here, we explain the crisis and implosion of the New Deal IR system as the

result of capital's regaining an upper hand from the 1970s on. The outlines of this story are well known: increasing globalization, transportation and communications technology that make it virtually costless to move production offshore, and the impetus to do so by the allure and opportunity of vast armies of cheap labor abroad. Missing from this standard story is a clear recognition that the U.S. decline in unionization and speed and extent of industrialization are unique—not found elsewhere in western and northern European nations or Japan. Our own view of American exceptionalism is that the formation and activism (agency, for short) of large industrial U.S. employers has been a pivotal component driving the United States' unique IR characteristics: the weakness of the labor left, the rapid destruction of unionization and collective bargaining in dominant industries, and the decline of U.S. labor standards and accompanying growth of inequality since the 1970s (Hillard & McIntyre, 2009a).

We begin with U.S. labor history from the 1880s to the 1920s. U.S. labor was arguably at that time as radical as European labor. But the ruthless use of force by U.S. capitalists and the weakness of the government in refereeing industrial relations (indeed, typically siding with employers by deploying troops) limited the success of the labor movement and cut off radicalism as a viable alternative to business unionism. Individual American employers had even greater relative incentive to repress labor because unionizing one company at a time put them at a competitive disadvantage; European unions had greater success in organizing whole industries (Jacoby, 1991; Wilentz, 1984).

American capitalist class exceptionalism thus traces its roots to the massive defeat of U.S. labor and popular organizations in the pre–New Deal period. Scholars comparing the labor histories of Europe and the United States conclude that U.S. employers went drastically farther and were singularly dedicated in their efforts to crush unions rather than accommodate them (Jacoby, 1991; Thelan, 2001). This exceptional commitment to crush labor was not fundamentally altered when the Great Depression and New Deal forced American capitalists, much against their will, to recognize unions' legal right to exist. As IR scholars cited above note, the New Deal period was one of vitality of non-union "welfare" employers, who sought to eliminate, not accommodate, the New Deal. Leading non-union employers, large and medium sized, led national efforts to weaken and repeal labor law and pioneered union-busting tactics (Jacoby, 1997). Modest victories turned into a collective rout when American capitalists succeeded in virtually wiping out private sector unions. The rout began in the early 1980s, when President Reagan fired 11,000 federally employed and unionized air traffic controllers on strike and permanently replaced them. As already unionized manufacturing industries contracted rapidly, depleting the numbers of existing workers with union representation, it became nearly impossible to unionize the growing service industries as employers learned, with the help of a huge industry of union-busting lawyers and consultants, that they could bend or violate labor law with impunity and prevent union campaigns from succeeding in three out of four cases, down from a more than 50% success rate in the 1940s–1960s (Freeman & Medoff, 1984; Logan, 2002). As noted below, private sector union representation in the United States has plummeted from more than 30% to single digits since 2000.

Capitalist class exceptionalism meant that free market or "neoliberal ideas" found fertile ground in the United States in the post-1970s. How exceptional is the United States? Recent work in comparative political economy places the United States in a group of "liberal market economies" that rely mostly on stock markets rather than institutional relationships in finance, give workers little in the way of employment protection, and have had high levels of inequality and relatively higher employment growth. Even within this group—all English-speaking countries—the United States stands out with much lower rates of union density, collective bargaining coverage, social spending, employment protection, family support, and poverty reduction through state redistribution, as well as much more income inequality (Pontusson, 2004). Thus, the loss of working-class economic and political strength has translated into worsened economic and social outcomes for the U.S. working class. As the Economic Policy Institute's excellent biannual publication, *The State of Working America,* has demonstrated, beginning in 1973 and accelerating in the 1980s and 1990s, median hourly wages have actually declined, low-wage jobs have proliferated, private employer–based retirement and health insurance has rapidly eroded, and annual hours worked have increased (the United States being the only advanced industrial nation that has not seen a decline of hundreds of hours annually worked), all at the same time that individual worker productivity has increased by approximately 80%. In Marxian terms, this would be an increase in exploitation sparked by the loss of working-class political and economic strength (Mishel, Bernstein, & Allegretto, 2005).

What made this possible is that the U.S. capitalist/employer class engaged in a process of "class formation" (i.e., becoming more united and stronger politically), and it built a political alliance with white working-class conservatives in rebellion against liberalism to place pro-business politicians and policies into government, symbolized by the "Reagan revolution," although it started under Democrats in the 1970s. In impressive fashion, when the opportunity for a comeback presented itself in the 1970s, U.S. capitalists built the institutions and attitudes necessary for action in their collective interest. This process of class formation included the growth of the Chamber of Commerce from 60,000 members in 1972 to more than a quarter million a year later; the movement of the National Association of Manufacturers to Washington, D.C., and the formation of the Business Roundtable, both also in 1972; and the establishment of the ultra-right Heritage

Foundation the next year, increased corporate backing for both the conservative American Enterprise Institute and the mainstream National Bureau of Economic Research, and the growing importance of right-wing foundations.

In sum, U.S. capital has maintained a long-run historical strength advantage over the U.S. working class. The New Deal era—1930s to 1970s—was a departure from this advantage because U.S. working-class strength was distinctly and unusually strong. Rapid erosion of the conditions that supported that strength (lack of international competition, a stable political coalition, etc.) and new opportunities allowed capital's ongoing efforts to defeat the New Deal IR system and dismantle it to succeed. This in turn left U.S. workers weak and their societal position ever lower, evidenced by the near destruction of private sector unionization, the consequent growth of lower waged work, and a steady weakening of the employer-based private health insurance and pension system (Klein, 2003), and a decline in U.S. labor standards (e.g., longer hours, low wages, lack of benefits).

## The Future of "Employment Relations" Study in the United States

Kaufman's influential 1993 book brought the "crisis" of the ILE/IR field into focus. It was clear that the world had changed. Specifically, the decline of unionized labor meant that the relevance of the field was called into question. Academic IR institutes were losing students and funding and were shutting down. For Kaufman, the only way out was to merge with the PM/HRM tradition, which many IR institutes did. With the decline of private sector unionization, its once influential role in making and implementing federal and state policy has severely diminished.

Despite this decline, IR scholars have continued to produce good scholarship. Emphasis on case studies, defining new issues—especially organizational/workplace redesign—and a frank recognition of previous mistakes, has produced a viable intellectual tradition. Scholarly practitioners have moved in two directions. One group has joined the broader labor studies scholarly community, which recognizes that society's progress depends on a strong labor movement. This has produced scholarship examining the use and outcome of power in union organizing. They have identified and scrutinized how employers have used legal and illegal tools to make union organizing nearly impossible, while shedding light on how clever service sector unions have overcome these barriers. Hotel workers, janitors, and low-wage health care workers have successfully organized themselves with union support through a "social movement" model. This model enlists workers' communities and community institutions in pressuring employers to recognize and bargain with workers (Bronfenbrenner, 2009; Clawson, 2003).

The other group of ILE/IR scholars has de-emphasized union strength (though quick to identify the positive role of unions in workplace productivity and implementing change) and turned to the question of workplace reform to improve the "competitiveness" of U.S. corporations. The chief recommendation here is to improve worker "voice" in management-defined "high-performance work systems" (HPWS; e.g., self-managed teams) (Applebaum, Bailey, Berg, & Kalleberg, 2000; Kochan & Osterman, 1994). In essence, these ILE/IR scholars fully joined the PM/HRM tradition in becoming management consultants, with the important caveats that the ILE scholars stressed the relevance if not superiority of implementing these reforms with the cooperation of unions and an emphasis on workers sharing in productivity gains. Regardless, this vein of work peaked in the 1990s. Continued rapid deindustrialization since 2000, widespread implementation of HPWS without gains sharing in service industries, and failure of signal examples such as Saturn Motors have sidelined the HPWS movement.

What of Marxian political economy and IR? Specifically, what are the opportunities for increased working-class political and economic strength in the twenty-first century? One component comes from a "de-centering of labor" (McIntyre & Hillard, 2007, 2009). That means recognizing that labor is performed not just in industrial and traditional service sector sites but also in the household ("domestic labor") and community. Some of this labor is paid a wage, but much is "unwaged," particularly domestic labor. Such labor has gone unrecognized by the IR field. The composition of what is considered work has also broadened; twenty-first-century labor is increasingly "immaterial" (producing services, not products) and comprises emotional, not physical, labor (caring for others; providing attention, support, reassurance). Gender is a related dimension. Domestic labor has been largely female. Women typically do waged work in emotional labor-intensive occupations. The mass entry of women into the waged workforce after 1970 has extended dramatically the problems of doing both a workplace and home shift. The "second shift" creates stress for households generally and especially for the two-shift workers who are predominantly women (Hochschild, 1989). These developments are an opportunity if, following the lead of a handful of progressive unions (e.g., the Harvard Clerical and Technical Workers), a new labor movement brings a decentered, gender-sensitive approach to union organizing, bargaining for more than the "hours, wages, and conditions" that defined the twentieth-century male "family wage." A gendered solidarity of women and men fighting for a transformation of both the household and the workplace represents the major opportunity of our time.

Finally, this model may currently exist only at the margins of our society but is not, at present, the kind of building and broad movement that labor saw a century ago when the labor question first arose. In addition, U.S. capital retains its

commitment to crush any progressive working-class movement. While the 2007–2009 financial and economic crisis, as well as Barack Obama's election, has notes of the 1934–1937 period, especially a discredited business class and free-market model, and while Obama is undoubtedly the most overtly pro-labor president in U.S. history by some measures, there is not a broad movement to carry out such an agenda. But if the crisis has taught us anything, it is that history takes surprising twists. And whatever shape it takes in the coming decades, class power will be a fundamental force in shaping that new history.

---

*Author's Note:* This chapter is in many ways a synopsis of a 20-year collaboration between the author and Richard McIntyre, Professor of Economics, University of Rhode Island. This work would also not be possible without the foundational work of Sanford Jacoby, Bruce Kaufman, and Thomas Kochan (see references below). The author thanks Professor McIntyre and Rhona Free for their helpful comments; any errors or omissions are solely the author's.

## References and Further Readings

Applebaum, E., Bailey, T., Berg, P., & Kalleberg, A. (2000). *Manufacturing advantage: Why high performance systems pay off.* Ithaca, NY: Cornell University Press, ILR Press.

Braverman, H. (1974). *Labor and monopoly capital.* New York: Monthly Review Press.

Brody, D. (1993). *Workers in industrial America: Essays in the twentieth century struggle* (2nd ed.). New York: Oxford University Press.

Bronfenbrenner, K. (2009). *No holds barred: The intensification of employer opposition to organizing* (EPI Briefing Paper #235). Retrieved from http://epi.3cdn.net/edc3b3dc172dd1094f_0ym6ii96d.pdf

Clawson, D. (2003). *The next upsurge: Labor and the new social movements.* Ithaca, NY: Cornell University Press, ILR Press.

Cowie, J. (1999). *Capital moves: RCA's 70-year quest for cheap labor.* Ithaca, NY: Cornell University Press, ILR Press.

Dubofsky, M. (1996). *Industrialism and the American worker: 1865–1920.* Wheeling, IL: Harlan Davidson.

Dunlop, J., Kerr, C., Harbison, F., & Myers, C. (1960). *Industrialism and industrial man: The problems of labor and management in economic growth.* Cambridge, MA: Harvard University Press.

Finegold, K., & Skocpol, T. (1984). State, party, and industry: From business recovery to Wagner Act in America's New Deal. In C. Bright & S. Harding (Eds.), *Statemaking and social movements: Essays in history and theory* (pp. 159–192). Ann Arbor: University of Michigan Press.

Fraser, S., & Gerstle, G. (Eds.). (1989). *The rise and fall of the New Deal order.* Princeton, NJ: Princeton University Press.

Freeman, R., & Medoff, J. (1984). *What do unions do?* New York: Basic Books.

Gordon, D., Edwards, R., & Reich. M. (1982). *Segmented work, divided workers.* New York: Cambridge University Press.

Green, J. (1998). *The world of the worker: Labor in twentieth century America.* Champaign: University of Illinois Press.

Harvey, D. (2005). *A brief history of neoliberalism.* New York: Cambridge University Press.

Hillard, M., & McIntyre, R. (1999). The crises of industrial relations as an academic discipline in the United States. *Historical Studies in Industrial Relations, 7,* 75–98.

Hillard, M., & McIntyre, R. (2009a). Historically contingent, institutionally specific: Class struggles, and American employer exceptionalism in the age of neoliberal globalization. In J. Goldstein & M. Hillard (Eds.), *Heterodox macroeconomics.* New York: Routledge.

Hillard, M., & McIntyre, R. (2009b). "IR experts" and the New Deal state: The diary of a defeated subsumed class. *Critical Sociology, 35,* 417–429.

Hochschild, A. (1989). *The second shift.* Berkeley: University of California Press.

Howe, S. (2002). *Empire: A very short introduction.* Oxford, UK: Oxford University Press.

Hyman, R. (1975). *Industrial relations: A Marxist introduction.* London: Macmillan.

Jacoby, S. (1991). American exceptionalism revisited: The importance of management. In S. Jacoby (Ed.), *Masters to managers: Historical and comparative perspectives on American employers* (pp. 173–200). New York: Columbia University Press.

Jacoby, S. (1997). *Modern manors: Welfare capitalism since the New Deal.* Princeton, NJ: Princeton University Press.

Kaufman, B. (1993). *The origins and evolution of the field of industrial relations in the United States.* Ithaca, NY: Cornell University Press, ILR Press.

Kaufman, B. (2004). *The global evolution of industrial relations.* Geneva, Switzerland: ILO Press.

Kerr, C., Dunlop, J., Harbison, F., & Myers, C. (1960). *Industrialism and industrial man.* Cambridge, MA: Harvard University Press.

Kessler-Harris, A. (2007). *Gendering labor history.* Urbana: University of Illinois Press.

Klein, J. (2003). *For all these rights.* Princeton, NJ: Princeton University Press.

Kochan, T., Katz, H., & McKersie, R. (1986). *The transformation of industrial relations.* Ithaca, NY: Cornell University Press, ILR Press.

Kochan, T., & Osterman, P. (1994). *The mutual gains enterprise.* Ithaca, NY: Cornell University Press, ILR Press.

Laurie, B. (1989). *Artisans into workers: Labor in nineteenth century America.* New York: Noonday Press.

Lichtenstein, N. (2002). *State of the union: A century of American labor.* Princeton, NJ: Princeton University Press.

Lichtenstein, N., & Harris, H. J. (Eds.). (1993). *Industrial democracy in America: The ambiguous promise.* Cambridge, UK: Cambridge University Press.

Logan, J. (2002). Consultants, lawyers, and the "union free" movement in the USA since the 1970s. *Industrial Relations Journal, 33,* 108–214.

Lynd, S. (Ed.). (1996). *We are all leaders: The alternative unionism of the early 1930s.* Urbana: University of Illinois Press.

Lynd, S. (1997). *Living inside our hope: A steadfast radical's thoughts on rebuilding movement.* Ithaca, NY: Cornell University Press, ILR Press.

Marx, K. (1977). *Capital* (Vol. I). New York: Vintage.

McIntyre, R. (2008). *Are workers rights human rights?* Ann Arbor: University of Michigan Press.

McIntyre, R., & Hillard, M. (2007). Decentering wage labor in contemporary capitalism. *Rethinking Marxism, 19,* 536–548.

McIntyre, R., & Hillard, M. (2009). A radical critique and alternative to U.S. industrial relations theory and practice. In F. Lee & J. Bekkan (Eds.), *Radical economics and labor* (pp. 133–142). New York: Routledge.

Minchin, T. (2001). *The color of work.* Chapel Hill: North Carolina University Press.

Mishel, L., Bernstein, J., & Allegretto, S. (2005). *The state of working America, 2004–2005.* Ithaca, NY: Cornell University Press.

Montgomery, D. (1980). *Workers' control in America: Studies in the history of work, technology, and labor struggle.* Cambridge, UK: Cambridge University Press.

Phillips-Fein, K. (2009). *Invisible hands.* New York: W. W. Norton.

Pontusson, J. (2004). *Inequality and prosperity: Social Europe vs. liberal America.* Ithaca, NY: Cornell University Press.

Ramirez, B. (1978). *When workers fight: The politics of industrial relations in the progressive era: 1898–1916.* Westport, CT: Greenwood.

Thelan, K. (2001). Varieties of labor politics in the developed democracies. In P. Hall & D. Soskice (Eds.), *Varieties of capitalism: The institutional foundations of comparative advantage* (pp. 71–103). New York: Oxford University Press.

Thompson, P., & Newsome, K. (2004). Labor process theory, work, and the employment relation. In B. Kaufman (Ed.), *Theoretical perspectives on work and the employment relationship* (pp. 133–162). Urbana, IL: Industrial Relations Research Association.

Tucker, R. (1978). *The Marx-Engels reader* (2nd ed.) New York: W. W. Norton.

Webb, B. (1901). *The case for factory acts.* London: Grant Richards.

Webb, S., & Webb, B. (1897). *Industrial democracy.* London: Longmans, Green.

Wilentz, S. (1984). Against exceptionalism: Class consciousness and the American labor movement. *International Labor and Working Class History, 26,* 1–26.

# PART II

## MICROECONOMICS

# 7

# Supply, Demand, and Equilibrium

Kevin C. Klein
*Illinois College*

The underlying foundation for much of the content for all other chapters in this reference manual is the economic concepts of supply and demand. In reading this chapter, you will begin to understand the basic concepts of supply and demand and how changes in the actions of buyers and sellers influence market prices. *Markets* are defined as any place where products or resources are exchanged. Every market has two sides: buyers and sellers. Buyers, or demanders, are those who purchase the product or resources. Sellers, or suppliers, are those who provide the products or resources for sale in the market. What motivates these market participants? Although many factors motivate buyers' and sellers' behavior, economists assume that the primary motivating factor is self-interest. Buyers are assumed to be motivated by their desire to improve overall satisfaction, or utility, in life. Sellers are assumed to be motivated by the desire to earn profits. The discussion below begins with the buyer side of the market, known as demand.

## Demand

Let us begin our discussion of demand by defining three concepts: demand ($D$), quantity demanded ($Q_d$), and the law of demand. *Demand* is defined as the amount of a product that buyers are willing and able to purchase at all prices. A consumer is said to demand a product if he or she is *both* willing and able to purchase a product. A consumer who is willing to purchase a product, but is unable to do so, is not considered to be part of the market demand because he or she will not actually purchase the product. Likewise, a consumer who is able but unwilling to buy a product is also not considered to be part of market demand. *Quantity demanded* is defined to be the amount of a product that buyers are willing and able to purchase at a *specific* price. To summarize, the difference between demand and quantity demanded for a product is that demand refers to the amount potential buyers are both willing and able to purchase at all prices, and quantity demanded refers to the amount potential buyers are willing and able to purchase at a specific price. Once these two terms are understood, it becomes possible to introduce the law of demand.

One of the most basic concepts in economics is the *law of demand.* The law of demand is simply an observation of a consumer's general response to changes in a product's price. As price decreases, consumers tend to be willing and able to purchase more of a product, and as price increases, consumers tend to be willing and able to purchase less of a product. To help visualize economic concepts, economists often try to illustrate the concepts using graphs. The law of demand is illustrated in Figure 7.1. The lines labeled $D_1$ and $D_2$ are referred to as demand curves. The lines are drawn here as linear functions to enhance the simplicity of the graph, but these lines could also be drawn as curved lines that are convex to the origin.

A *change in demand* is illustrated by a shift in the location of the entire curve. In Figure 7.1, demand is said to increase if demand changes from $D_1$ to $D_2$. An increase in demand means that consumers are willing and able to purchase more at every price. For example, the movement from point *A* to *C* represents an increase in demand because at point *A,* consumers are willing to buy $Q_1$, but $Q_2$ at point *C.* This increase in demand occurs even though the price remains unchanged at $P_1$. A decrease in demand means the consumers are willing and able to purchase less at every price and is illustrated by a shift in demand from $D_2$ to $D_1$. A *change in quantity demanded* is illustrated by

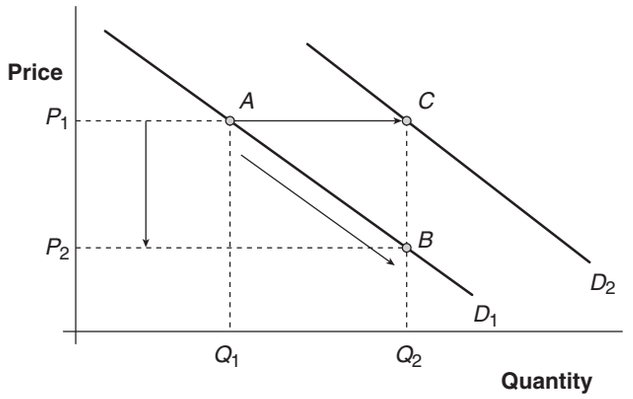


**Figure 7.1** Changes in Demand Versus Changes in Quantity Demanded

a move along the curve. In Figure 7.1, an increase in quantity demanded is illustrated by a movement from point *A* to *B*. It is important to note that the increase in consumption from $Q_1$ to $Q_2$ occurs because price has decreased from $P_1$ to $P_2$. A decrease in quantity demanded is illustrated by a movement from point *B* to *A*.

## Determinants of Demand

Many factors change a buyer's willingness or ability to purchase goods and services. These factors are known as determinants of demand. Consider the market for gasoline and what factors might influence a buyer's willingness or ability to purchase gasoline. To determine the impact of changes in each of the factors listed below, economists usually invoke the ceteris paribus assumption. *Ceteris paribus* is a Latin term usually interpreted to mean "all other factors remaining unchanged." This assumption is made to isolate the effects of a change in the factor under consideration.

### Tastes

Tastes refer to whether buyers like or dislike a product. If buyers' tastes change in favor of a product, ceteris paribus, demand increases. In the market for gasoline, when buyers' tastes change in favor of a product, the buyer is willing and able to purchase more gas at every price. Remember that this is illustrated as a movement from $D_1$ to $D_2$ in Figure 7.1. An increase in taste for gasoline would occur, for example, when consumers' tastes change toward larger vehicles such as sport utility vehicles, minivans, trucks, and high-performance sport cars. An increase in demand would also occur as consumers move to houses or apartments that are larger driving distances from their places of employment. If consumers' tastes change toward smaller, more fuel-efficient cars, or consumers prefer to live closer to work, the demand for gasoline will decrease. This is illustrated as a shift of the demand curve from $D_2$ to $D_1$.

### Income

Income refers to the amount of money consumers have available to spend on goods or services. This directly affects consumers' ability to purchase a good or service. The relationship between consumers' purchases and income can be broken into two categories.

1. *Normal goods:* Normal goods are any good or service for which consumers tend to purchase more as their income increases, ceteris paribus. Gasoline is generally considered to be a normal good. As consumers' incomes increase, they tend to purchase more gasoline for a variety of reasons. For example, some people who cannot afford a car at lower income levels may be able to purchase a car at higher income levels, and thus their demand for gasoline increases. Many other income-related reasons exist for increases in the demand for gasoline. Two other examples include more frequent and longer vacations and more cars per household as income increases. Each of these examples results in a shift from D1 to D2 in Figure 7.1.

2. *Inferior goods:* Inferior goods are any good or service for which consumers tend to purchase less as their income increases. The opposite is also true. That is, as income decreases, consumption of inferior goods tends to increase. For example, during recessions, economists expect to see consumers buying more store-brand groceries instead of name-brand products. Under this scenario, the store-brand products would be considered inferior goods. An increase in income will cause the demand for an inferior good to change from $D_2$ to $D_1$.

### Prices of Related Goods

Changes in the prices and availability of related goods will also affect the demand for a product. The relationship to related goods is broken into two categories.

1. *Substitutes:* Substitutes are items consumed in place of other goods. In the case of gasoline, several substitutes are available. If, for example, gasoline is being used to fuel a car, then consumers might purchase an electric car, thus making electricity a substitute for gasoline. Another alternative to gasoline is an ethanol blend. For example, E85 is 85% ethanol and 15% gasoline. If the price of gasoline increases, ceteris paribus, consumers will want to purchase less gasoline, and thus the demand for its substitute will increase. Yet another alternative to gasoline is diesel fuel. An increase in the price of gasoline, ceteris paribus, would cause an increase in the demand for diesel fuel as consumers switch to diesel-powered cars and trucks to avoid the rising price of gasoline.

2. *Complements:* Complements are items consumed together. In the case of gasoline, anything consumed with gasoline is its complement. For example, automobiles and trucks are complements to gasoline. Based on the law of demand, if the price of gasoline were to increase, ceteris paribus, economists would expect consumers to purchase

less gasoline. On the demand curve in Figure 7.1, this is represented as a *move along the curve* from point *B* to *A*. The decrease in purchases of gasoline would then result in fewer miles driven. Driving fewer miles results in less wear and tear on automobiles, resulting in a decrease in the demand for cars and trucks. In the market for cars and trucks, the decrease in demand resulting from an increase in the price of gasoline is represented as a movement from point *C* to *A* as the consumer moves from $D_2$ to $D_1$.

### Expectations

Consumer demand is often affected by what the consumers anticipate will happen in the future. For example, in the case of gasoline, if consumers expect the price of gasoline to continue to rise, then in the short run, consumers will quickly go out and purchase more gasoline even thought the price has not yet changed. Consequently, demand increases, shifting demand from $D_1$ to $D_2$ or a movement from point *A* to *C* in Figure 7.1. In the long run, in anticipation of long-term increases in the price of gasoline, consumers may purchase more fuel-efficient cars and trucks. These expectations result in a decrease in the demand even though the price of gas has not yet changed. Graphically, this is represented as a movement from point *C* to *A* in Figure 7.1.

### Number of Consumers

The number of people who are willing and able to purchase a product directly affects the demand for a product. In the market for gasoline, as the number of consumers increases, ceteris paribus, the demand for gasoline increases. As the number of consumers decreases, the demand for gasoline decreases.

### Government Regulations

Regulations can have a variety of effects on the demand for a product. Regulations can either increase or decrease the demand for a product depending on the nature of the regulation. For example, if the government were to increase the Corporate Average Fuel Efficiency standards for automobiles, the demand for gasoline would go down as the nation's car fleet becomes more fuel efficient. This change in standards would result in shifting the demand for gasoline from $D_2$ to $D_1$ in Figure 7.1.

It is useful to remember that any change of the determinants of demand would cause a shift of the entire demand curve, not a movement along the curve. The demand curve shifts to the right when demand increases and to the left when demand decreases.

## Supply

The other group of participants in markets is the sellers. When discussing supply, there are again three concepts to define and discuss: supply, quantity supplied, and the law of supply.

*Supply* is defined to be the amount of a product sellers are both willing and able to provide to the market at all prices. Key to this definition is that sellers have to be both willing and able to provide the product to the market. A seller who is willing to sell the product but unable to do so is not considered to be part of supply because he or she will not actually provide the product to the market. Likewise, sellers who are able but unwilling are also not considered to be part of market supply. *Quantity supplied* is defined to be the amount of a product that sellers are both willing and able to provide to the market at a specific price. The difference in these two concepts is similar to the difference between demand and quantity demanded. That is, supply refers to the entire supply relationship, while quantity supplied refers to a single price and quantity combination. Once these two concepts are understood, it becomes possible to state the *law of supply*. According to the law of supply, as price increases, sellers increase the quantity that they are willing and able to provide to the market. If price decreases, sellers decrease the quantity that they are willing and able to provide to the market. The lines labeled $S_1$ and $S_2$ in Figure 7.2 are referred to as supply curves. The lines are drawn here as linear functions to enhance the simplicity of the graph, but these lines could also be drawn as curved lines that are convex to the origin.

A *change in supply* is illustrated by a shift in the location of the entire curve. In Figure 7.2, supply is said to increase if supply changes from $S_1$ to $S_2$. An increase in supply means that sellers are willing and able to sell more at every price. For example, the movement from point *E* to *G* represents an increase in supply because at point *E,* sellers are willing to sell $Q_3$, but $Q_4$ at point *G*. This increase in willingness to sell occurs even though the price remains unchanged at $P_1$. A decrease in supply means the sellers are willing and able to sell less at every price and is illustrated by a shift in demand from $S_2$ to $S_1$. A *change in quantity supplied* is illustrated by a move along the curve. In Figure 7.2, an increase in quantity supplied is illustrated by a movement from point *F* to *G*. It is important to note that the increase in the willingness to sell from $Q_3$ to $Q_4$ occurs because price has increased from $P_2$ to $P_1$. A decrease in quantity supplied is illustrated by a movement from point *G* to *F*.

## Determinants of Supply

As we saw with demand, a number of factors influence a seller's willingness and ability to provide a good or service to the market. These factors are known as *determinants of supply*. Consider again the market for gasoline. What factors might influence a seller's willingness or ability to sell gasoline?

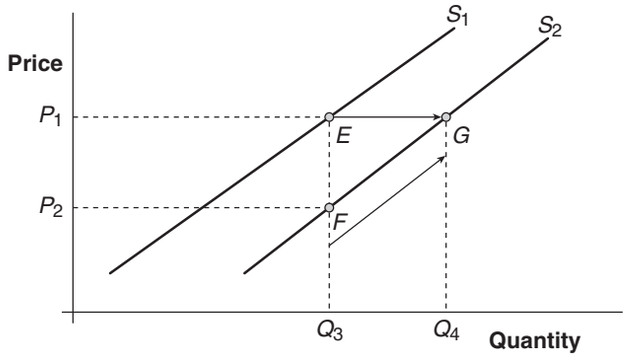


**Figure 7.2** Change in Supply Versus Change in Quantity Supplied

## Factor Costs

Factor costs refer to the amount of money paid for inputs in the production process. If factor costs increase, then supply will decrease, represented by a movement from $S_2$ to $S_1$ in Figure 7.2. For example, one of the factors of production used to produce gasoline is crude oil. If the cost of crude oil increases, this raises the cost of producing gasoline. As a result, the amount of gasoline available at every price will decrease. Consider for a moment that suppliers of gasoline are willing to sell $Q_3$ gasoline at a price of $P_2$. If the cost of producing gasoline increases, due to an increase in the cost of crude oil, then sellers would expect a higher price, perhaps $P_1$, to be willing and able to sell the same quantity. Because price must rise for suppliers to keep the quantity supplied unchanged, supply must decrease if price remains unchanged. This scenario describes a decrease in supply and is illustrated by a shift from $S_2$ to $S_1$ in Figure 7.2. If factor costs decrease, then supply will increase, represented by a movement from $S_1$ to $S_2$ in Figure 7.2.

## Technology

The quantity and quality of technology available for use in the production process affects the supply. In general, improvements in technology lower the cost of producing most goods and services. It then follows that new cost-lowering technologies result in increasing supply. This is represented by a movement from $S_1$ to $S_2$ in Figure 7.2.

## Price of Related Goods in Production

Changes in the prices and availability of related goods *in production* will affect supply. These relationships are broken into two categories.

1. *Substitutes:* Substitutes in production are items that can be produced or sold in place of other items. For example, an oil refinery can switch its production between different octane levels of gasoline depending on which octane level has a higher demand and offers higher profits. Let us say a refinery is selling 87 octane gasoline but the profits for 93 octane gasoline increase. The refiner would then want to increase its refining of 93 octane gasoline. Because the refinery has limited refining capacity, it must choose between refining more 93 octane gasoline and maintaining its current refining allocation between 87 and 93 octane gasoline. If the refinery chooses to produce more 93 octane gasoline, then the supply of 87 octane gasoline will decrease, represented by a movement from $S_2$ to $S_1$ in Figure 7.2. Note that the price of 87 octane gasoline is assumed to have remained unchanged, yet the refinery decreases its production of 87 octane gasoline. It did so because the profitability of a substitute *in production* increased.

2. *Complements:* Complements in production are products that are produced in conjunction with another product. For example, when refining crude oil for gasoline, a by-product of the refining process is liquefied petroleum gas, also known as propane. If, for example, the price of gasoline increases and the refinery wants to increase the amount of gasoline it refines, this will necessarily result in an increase in the supply of propane because gasoline and propane are complements in production. To produce more gasoline requires that more propane is also produced during the refining process. In the gasoline market described here, there has been an increase in quantity supplied, a movement from point $F$ to $G$, but in the propane market, there would be a shift in the entire supply curve from $S_1$ to $S_2$.

## Expectations

A seller's willingness and ability to provide products to the market are affected by his or her forecasts for the future. If, for example, managers of gasoline refineries expect the price of gasoline to increase, they may withhold gasoline from the market so they can sell their gasoline at higher prices. As the refineries withhold their gas from the market in anticipation of the future higher prices, the supply curve decreases, shown as a shift in supply from $S_2$ to $S_1$ or a movement from point $G$ to $E$ in Figure 7.2. In the long run, sellers of gasoline might anticipate that the population will increase and increase future sales of gasoline. In response, they may build new refineries in anticipation of the increased demand. This forecast, or expectation, of future events results in an increase in supply of gasoline, shown as a shift from $S_1$ to $S_2$ in Figure 7.2.

## Number of Sellers

The number of sellers in the market directly affects the amount of the product available for sale at every price. An increase in the number of sellers increases supply from $S_1$ to $S_2$ in Figure 7.2. A decrease in the number of sellers will decrease supply from $S_2$ to $S_1$.

### Government Regulations

Regulations can have a variety of effects on the supply of a product. Regulations can either increase or decrease the supply of a product depending on the nature of the regulation. For example, in the market for gasoline, the government could impose stricter pollution standards for oil refineries. As a result, controlling pollution from the refineries would increase and the cost of refining gasoline would also increase. As noted earlier, increases in costs result in decreasing supply. In Figure 7.2, this is represented as a movement from point $G$ to $E$ or from $S_2$ to $S_1$.

It is useful to remember that any change of the determinants of supply would cause a shift of the entire supply curve, not a movement along the curve. The supply curve shifts to the right when supply increases and to the left when supply decreases.

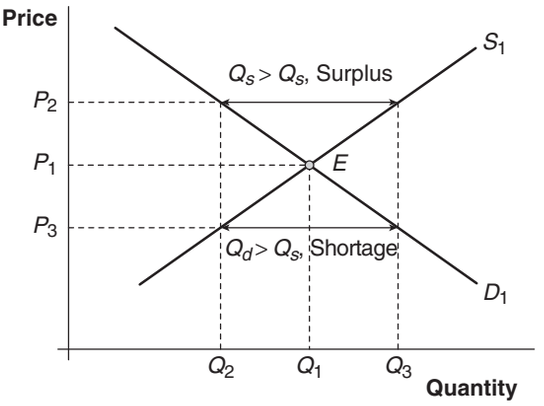


**Figure 7.3** Equilibrium

NOTES: Equilibrium occurs when $Q_s = Q_d$ at a common price. If market price is above the equilibrium price, a surplus forms. If the market price is below the equilibrium price, a shortage forms.

## Equilibrium

All markets are characterized by market participants making decisions to improve their own self-interest. As noted earlier, buyers are assumed to be motivated primarily by the desire to increase their personal satisfaction in life. Sellers are assumed to be motivated by the desire to make the largest profits possible. This interaction between buyers and sellers results in equilibrium in the market. *Equilibrium* is defined to be the point at which $Q_s = Q_d$ at a common price. If the market price is not at equilibrium, market forces drive the market toward equilibrium. If the market price is at equilibrium, there exists no market pressure to move to some other level. To better understand this concept, consider Figure 7.3.

The market represented in Figure 7.3 is in equilibrium at point $E$ because $Q_s = Q_d$ at the common price of $P_1$. To better understand the concept of equilibrium, it is useful to examine the market at other prices. First, consider a market price above the equilibrium value. At $P_2$, the quantity supplied, $Q_3$, is larger than the quantity demanded, $Q_2$. At this price, sellers are willing to provide more to the market than consumers are willing to buy. This situation is defined as a *surplus.* To combat their rising inventories, sellers begin to lower their prices, attempting to entice consumers to buy more. In addition, because of the lower price, sellers decrease the amount they offer for sale. As this process unfolds, there is an increase in $Q_d$ and a decrease in $Q_s$ as the market moves to the equilibrium at point $E$.

Now consider a market price below the equilibrium value. At the price $P_3$, the quantity demanded, $Q_3$, exceeds the quantity supplied, $Q_2$, resulting in a shortage in the market. In response to the shortage, consumers begin to bid up the price. To understand this process, consider the process of an auction. An auction usually begins with the number of willing and able buyers exceeding the amount of product available. The role of the auctioneer is to bid the price up until $Q_s = Q_d$. As the price increases, consumers begin to drop out of the bidding process. In the market represented in Figure 7.3, consumers begin bidding the price up and, as a result, some consumers begin to drop out of the market and $Q_d$ begins to fall. As the price begins to increase, sellers who are willing and able respond by increasing the amount they provide to the market, and $Q_s$ begins to increase. These changes in $Q_s$ and $Q_d$ continue until the market reaches equilibrium at a price of $P_1$ and a quantity of $Q_1$.

Ultimately, markets tend to move toward equilibrium. When market prices are too high, surpluses exist, and market forces drive the prices down to the equilibrium level. When market prices are too low, shortages exist, and market forces drive up the prices to the equilibrium price. At equilibrium, there exists no market pressure for change, and thus the price tends to stay at the equilibrium until something in the economy changes, resulting in a change in either supply or demand.

### The Algebra of Demand and Supply

The law of demand suggests that an inverse mathematical relationship exists between price and quantity demanded. The law of supply suggests that a similar but positive relationship exists between price and quantity supplied. An example of these relationships is shown in Equations 1 and 2. Assume the following hypothetical functions represent the relationship between price and quantity for gasoline:

$$\text{Demand: } P = 1{,}210 - .05Q_d. \quad (1)$$

$$\text{Supply: } P = 10 + .01Q_s. \quad (2)$$

Remember that at equilibrium, $Q_s = Q_d$ at a common price. Thus, at equilibrium, these two equations are equal.

As a result, it is possible to solve for the equilibrium price and quantity in this hypothetical model. Setting these two equations equal results in

$$1{,}210 - .05Q_d = 10 + .01Q_s. \quad (3)$$

At equilibrium, $Q_s = Q_d = Q$, and thus we can substitute $Q$ for $Q_s$ and $Q_d$. Solving for $Q$ gives us the equilibrium quantity in this market.

$$1{,}200 = .06Q, \text{ thus } Q = 20{,}000. \quad (4)$$

Substituting $Q = 20{,}000$ into either Equation 1 or 2 results in the equilibrium price.

$$P = 1{,}210 - .05Q_d = 1{,}210 - .05(20{,}000) = 210. \quad (5)$$

## Changing Equilibrium Prices and Quantity

Although markets tend to move toward equilibrium over time, should any of the determinants of supply or demand change, the result will be a tendency toward a different equilibrium.

### *The Effects of a Change in Demand*

The impact on equilibrium price and quantity of a change in demand is illustrated in Figure 7.4.

Figure 7.4 represents the market for gasoline. Assume the market is initially in equilibrium at point $E_1$. Further assume that national income statistics indicate an increase in household income, ceteris paribus. What impact would this increase in income have on the equilibrium price and quantity in the market for gasoline? At the original equilibrium price of $P_1$, consumers are now willing and able to purchase more gasoline even though the price of gasoline initially remains unchanged. In Figure 7.4, this is represented by a movement from point $E_1$ to $A$ and an increase in quantity from $Q_1$ to $Q_3$. At the initial price of $P_1$, this increase in demand causes a shortage of $(Q_3 - Q_1)$. As shown earlier, shortages result in increases in price. The increasing price causes a decrease in quantity demanded along the new demand curve, $D_2$, and an increase in quantity supplied along the supply curve. When all of these factors are combined, the market moves toward a higher equilibrium price and quantity at $P_2$ and $Q_2$.

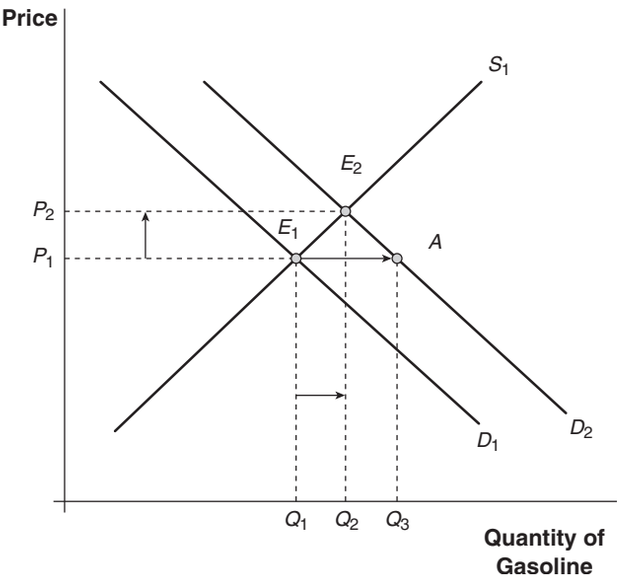


**Figure 7.4** Effects of a Change in Demand

NOTE: An increase in demand results in an increase in equilibrium price and an increase in quantity supplied from $Q_1$ to $Q_2$.

### *The Algebra of a Change in Demand*

Algebraically, a change in demand will change the demand function. For example, an increase in demand might be represented in our earlier hypothetical market for gasoline as

$$D\text{: } P = 1{,}510 - .05Q_d. \quad (6)$$

Given the original supply function of $P = 10 + .01Q_s$, solving the system of equations yields a new equilibrium price and quantity combination of $P = 260$, $Q = 25{,}000$.

This new equilibrium value verifies our graphical predictions. That is, an increase in demand should result in an increase in equilibrium price and quantity. In this case, the equilibrium price increased from 210 to 260, and the equilibrium quantity increased from 20,000 to 25,000 during the given time period.

### *The Effects of a Change in Supply*

Assume the number of sellers increases, ceteris paribus (see Figure 7.5).

The shift from $S_1$ to $S_2$ represents the increase in supply resulting from an increase in the amount of sellers who are willing and able to provide gasoline at every price. At the original equilibrium price $P_1$, this increase in supply results in a surplus equal to $(Q_3 - Q_1)$. As noted earlier, surpluses result in downward pressure on equilibrium price. As sellers reduce price and production to reduce the surplus in the market, buyers increase their quantity demanded in response to the lower price. The combination of these forces results in a decrease in equilibrium price from $P_1$ to $P_2$ and an increase in equilibrium quantity from $Q_1$ to $Q_2$.

### *The Algebra of a Change in Supply*

Algebraically, a change in supply is demonstrated by a change in the supply function. For example, an increase in supply might be represented in our earlier hypothetical market for gasoline as

$$\text{Supply: } P = 4 + .01Q_s. \quad (7)$$

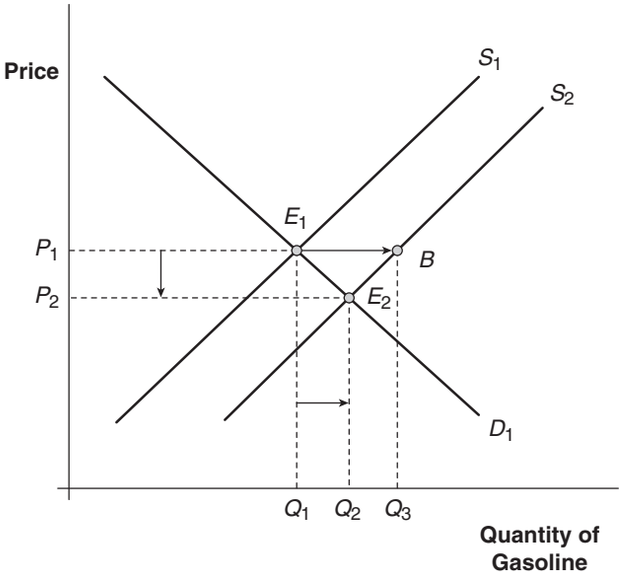


**Figure 7.5** The Effects of a Change in Supply

NOTE: An increase in supply results in a decrease in equilibrium price and an increase in quantity demanded from $Q_1$ to $Q_2$.

Given the new supply function in Equation 7 and the new demand function presented in Equation 6, solving the system of equations yields a new equilibrium price and quantity combination of $P = 255$, $Q = 25{,}100$. This change in equilibrium price and quantity verifies the graphical predictions that an increase in supply will result in a decrease in equilibrium price and an increase in equilibrium quantity. In this case, the equilibrium price has decreased from 260 to 255, and the equilibrium quantity has increased from 25,000 to 25,100.

### *The Combined Effects of Changing Supply and Demand*

If both supply and demand change simultaneously, the impacts on price and quantity may not be as certain as described in the two previous examples. Remember that an increase in supply results in a decrease in equilibrium price and an increase in equilibrium quantity. An increase in demand results in an increase in equilibrium price and an increase in equilibrium quantity. What would be the impact on equilibrium price and quantity if supply and demand both increased simultaneously? In such cases, the individual effects are combined. Because the increases in both supply and demand result in increases in the equilibrium quantities, the combined result will be an increase in equilibrium quantity. However, the result on equilibrium price is uncertain. When both supply and demand increase simultaneously, the downward pressure on price, due to increasing supply, is opposite of the upward pressure on price due to increasing demand. In such cases, it becomes impossible to predict the direction of change without additional detail about the magnitude of the changes in supply and demand. As a result, we can only conclude the direction of change is uncertain. In situations where both supply and demand change simultaneously, the direction of change in either equilibrium price or equilibrium quantity—but not both—will be uncertain unless specific information is available about the magnitude of the change. For example, when comparing Equations 1 and 2 to Equations 6 and 7, we can predict with certainty the changes in equilibrium price and quantity because the estimates of demand and supply, shown in the equations, provide the magnitude of change. If statistical data are available that allow for an accurate estimate of the supply and demand functions, the ambiguity in the direction of change is eliminated. For more information on how to develop these statistical estimates, read the econometrics chapter of this handbook, Chapter 5.

## Price Controls: Ceilings and Floors

The market process described above assumes that market participants are each making decision based on their own self-interest. Periodically, a market's equilibrium price and quantity are viewed by some market participants as undesirable. When this happens, participants on one side of the market may lobby for government intervention. If, for example, consumers believe the market equilibrium price is too high, they may lobby for governments to impose a price ceiling. A *price ceiling* is a legally established maximum price intended to lower the price below the market equilibrium price. If government agrees to establish an effective price ceiling, the ceiling will be established below the equilibrium price, as illustrated in Figure 7.6.

Recall that when market price is below the equilibrium value, the resulting shortages cause prices to increase toward the equilibrium. However, when a price ceiling is established, the ceiling price is a maximum price, and sellers cannot legally charge a higher price. As a result, the shortages created by the price ceiling are persistent and an expected long-term outcome resulting from the price ceiling. Price ceilings thus have interesting policy implications. In choosing to protect some consumers from the unwanted market price, the government is also choosing to deny other consumers access to the product as a result of the shortage. In general, economists do not recommend implementation of price ceilings because of the resulting persistent shortages. Perhaps the most famous example of price ceilings is rent controls in New York City. Although some people are able to acquire rent-controlled housing, there is a persistent shortage of housing within the city.

If producers believe the market price is too low, they may lobby for governments to impose a price floor. A *price floor,* also known as a price support, is a legally established minimum price. If the intent of establishing a price floor is to raise the price above the market equilibrium price, the floor will be established above the equilibrium price. As shown in Figure 7.6, a price floor set above the equilibrium price will result in a surplus. The price floor sends a signal to the sellers to increase their production, and for consumers, the market signal sent is to decrease their

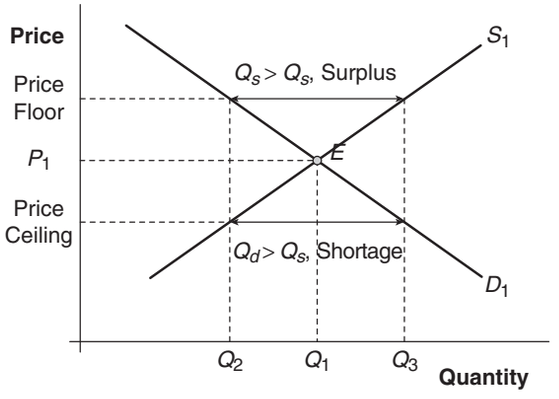


**Figure 7.6** The Effects of Price Ceilings and Floors

NOTES: Effective price ceilings result in market shortages. Effective price floors result in market surpluses.

consumption compared to the equilibrium level. The effects of a price floor are shown in Figure 7.6. The price floor encourages sellers to offer a quantity $Q_3$ for sale but also encourages consumers to buy a quantity of $Q_2$. The result is a surplus of $(Q_3 - Q_2)$. When a market price is above the equilibrium value, the resulting surpluses are normally eliminated by declining prices. However, if a price floor is established, the price floor prevents the price from falling to the equilibrium level. As a result, the surpluses created are persistent and an expected long-term outcome resulting from the price floor. Again, this result has interesting policy implications. In choosing to protect some sellers from the unwanted market price, the government is also choosing to create an unwanted surplus. For example, agricultural policies in many countries are often designed to establish price floors above the equilibrium prices. The price floors send a signal to the market participants that result in surpluses. Consequently, governments must enact other policies to deal with the problems created by the surpluses such as long-term storage and the pressures on market prices of accumulating surpluses. In general, economists do not recommend implementation of price floors. Although the price received by the producer is higher, the resulting surpluses create a new problem. To respond to the new problem, government must enact an increasing level of other policies to deal with the persistent and growing surpluses. More information on agricultural economics can be found in Chapter 58 of this handbook.

*The Algebra of Price Ceilings and Floors*

Applying algebra to the concepts of price ceilings and floors is a simple extension of the algebraic model shown earlier. Let us return to the original equations with an equilibrium price and quantity of $P = 210$ and $Q = 20{,}000$. For a price floor to accomplish its intended goal, it must be established above the equilibrium price. Assume that a price floor, $P_F$, is established at $P_F = 250$. What impact will this have on the market? With the minimum price established at $P_F = 250$, consumers are willing and able to purchase 19,200 units of gasoline based on Equation 1. At that same price, sellers are willing and able to sell 24,000 units of gasoline based on Equation 2. Thus, the price ceiling established at $P_F = 250$ results in a surplus in the market of 4,800 units.

For a price ceiling to accomplish its intended goal, it must be established below the equilibrium price. Assume that a price ceiling, $P_C$, is established at $P_C = 190$. What impact will this have on the market? With a maximum price established at $P_C = 190$, consumers are willing and able to purchase 20,400 units of gasoline, based on Equation 1, and sellers are willing and able to sell 18,000 units of gasoline, based on Equation 2, resulting in a shortage of 2,400 units of gasoline in the market.

## Case Studies

Examining a few case studies can help us better understand the dynamics of supply and demand.

### Case 1: The Market for Gasoline

The first case study is for gasoline in the Phoenix, Arizona, market in 2003. According to the Federal Trade Commission (FTC, 2005), in August 2003, the average price of gasoline in the Phoenix market rose from $1.52 per gallon during the first week of August to $2.11 by the third week of August. By the end of September, the price fell to $1.80 per gallon. What could have caused these abrupt changes in prices? Although collusion by suppliers was considered a possible cause of the price spike, the FTC concluded in its 2005 report that

> the price spike was caused by a pipeline rupture on July 30, and the failure of temporary repairs, which had reduced the volume of gasoline supplies to Phoenix by 30 percent from August 8 through August 23. Arizona has no refineries. It obtains gasoline primarily through two pipelines, one traveling from west Texas and the other from the West Coast. The rupture closed the portion of the Texas line between Tucson and Phoenix. (p. 5)

This conclusion is consistent with the equilibrium analysis shown above in Figure 7.5. The pipeline rupture resulted in a sudden decrease in the supply of gasoline to the Phoenix market. A decrease in supply results in a shortage of gasoline at the original equilibrium and a subsequent increase in price as the market moves to a new equilibrium. This can be represented as a shift from $S_2$ to $S_1$ and a movement from $E_2$ to $E_1$ in Figure 7.5. The report continues to state,

> The shortage of gasoline supplies in Phoenix caused gasoline prices to increase sharply. To obtain additional supply, Phoenix gas stations had to pay higher prices to West Coast refineries

> than West Coast gas stations were paying. West Coast refineries responded by selling more of their supplies to the Phoenix market. (FTC, 2005, p. 5)

The increase in quantity supplied by the West Coast refineries, along the new supply curve $S_1$, is consistent with the theory of supply stated above.

How did Phoenix consumers respond to the higher prices? Based on the law of demand, when prices rise, the quantity demanded should decline in response to higher prices. However, the degree to which consumers respond to higher prices depends on a number of factors, including the amount of substitutes available, the length of time involved, and the product price relative to the consumers' income or budget. The FTC (2005) commented on the consumer response by stating,

> Phoenix consumers did not respond to significantly increased gasoline prices with substantial reductions in the amount of gasoline they purchased. In theory, to prevent a gasoline price hike, Phoenix consumers could have reduced their gasoline purchases by 30 percent. Without price increases, however, consumers do not have incentives to change the amount of gasoline they buy. Moreover, even with price increases, most consumers do not respond to short-term supply disruptions such as a pipeline break by making the types of major changes—the car they drive, their driving habits, where they live, or where they work—that could substantially reduce the amount of gasoline they consume. . . . Empirical studies indicate that consumers do not easily find substitutes for gasoline, and that prices must increase significantly to cause even a relatively small decrease in the quantity of gasoline consumers want. In the short run, a gasoline price increase of 10 percent would reduce consumer demand by just 2 percent, according to these studies. This suggests that gasoline prices in Phoenix would have had to increase by a large amount to reduce the quantity of consumers' purchases by 30 percent, the amount of lost supply. (p. 5)

Readers who are interested in more information about consumer responses to price changes should read Chapter 9, Demand Elasticities, in this handbook.

## Case 2: The Market for Corn

A second interesting question to study is whether the price of corn is affected by changes in the market for biofuels. One major type of biofuel is ethanol. Ethanol is a combustible fuel that can be distilled from corn and, when mixed with gasoline, used in internal combustion engines. Rising world energy prices, combined with concerns about the potential impacts of carbon emissions on global climate change, have resulted in renewed interest in ethanol as a renewable alternative fuel for automobiles. Proponents of ethanol use claim not only that ethanol is a renewable form of energy but also that blending ethanol with gasoline will also reduce the total amount of greenhouse gas emissions. That claim is supported in an April 2009 Congressional Budget Office (CBO) report titled *The Impact of Ethanol Use on Food Prices and Greenhouse-Gas Emissions.* In this report, the CBO states that research conducted by Argonne National Laboratory suggests that, in the short run, production and consumption of ethanol will create up to 20% less greenhouse gases than will the equivalent production and consumption of gasoline (CBO, 2009, p. 10). The long-run impact of ethanol on greenhouse emission is not as clear. Regardless of the long-run impact on greenhouse emission, a series of legislative acts involving mandated use of biofuels have had a significant impact on the market for corn. Most recently, the Energy Independence and Security Act of 2007 expanded mandated use of biofuels through 2022 and will ultimately result in the production of 15 million gallons of corn-based ethanol (CBO, 2009, p. 13).

The analytical question for this case is, what impact will the use of corn for ethanol production have on the price of corn? Simple supply-and-demand analysis can be used to help determine the answer to this question. Increases in the demand for any product should, ceteris paribus, result in increases in the price. In the case of corn-based ethanol, the increased demand for corn used to make ethanol has resulted in an increase in the price of corn. The CBO (2009) report states,

> The upswing in the demand for corn to be used in producing domestic ethanol raised the commodity's price, CBO estimates, by between 50 cents and 80 cents per bushel between April 2007 and April 2008. That range is equivalent to between 28 percent and 47 percent of the increase in the price of corn, which rose from $3.39 per bushel to $5.14 per bushel during the same period. . . . As the mandated use of biofuels rises over time, increased production of ethanol and biodiesel will probably continue to push up prices for corn and soybeans. According to an estimate by the Department of Agriculture, 3.7 billion bushels of corn will be used to produce ethanol during the 2008–2009 marketing year. That estimate represents an increase of about 0.7 billion bushels over the total for the previous marketing year, which could increase the price of corn by 10 percent to 17 percent over the 2008–2009 period, all else being equal. In the long run, upward pressure on prices caused by increasing ethanol production may be alleviated by planting additional acres in corn and soybeans, increasing crop yields per acre in the United States and abroad, and improving the technologies used at refineries to allow more ethanol to be produced from each bushel of corn. (p. 18)

The CBO information can be analyzed using supply-and-demand analysis. The biofuel mandates by the Energy Independence and Security Act of 2007 will continue to increase the demand for corn. An increase in demand, ceteris paribus, will result in an increase in equilibrium price and a subsequent increase in quantity supplied, as demonstrated in Figure 7.4. For example, the CBO (2009) analysis states that in 2007, corn prices increased between 50 and 80 cents due to the increased demand for corn resulting from the production of ethanol. The biofuel mandates, which call for an increase in the use of biofuels, should thus cause a further increase in the demand for corn, ceteris paribus, and result in a continued increase in corn prices.

The long-run impact of ethanol production on corn prices is more difficult to determine because, over time, markets respond to changing prices in a variety of ways. For example, in anticipation of higher corn prices and profits, farmers are likely to plant more acreage in corn. Seed corn companies are likely to develop varieties of corn that can be planted in regions of the nation that are not traditional corn-growing regions. Growers worldwide are likely to convert acreage, including clear cutting forest areas, which have not been traditionally planted in corn. Rising corn prices will cause ethanol producers to search for alternatives to corn to make ethanol and, consequently, decreasing demand for corn. The resulting long-term change in supply and demand for corn results in an ambiguous change in the price of corn.

## Conclusion

Markets for all goods and services have two sides: buyers and sellers. The interaction of the buyers and sellers, all acting in their own self-interest, establishes an equilibrium price and quantity. While these equilibriums tend to be stable, changes in the buyers' or sellers' willingness and ability to participate in the market result in changing equilibrium prices and quantities. Sometimes, either consumers or producers are dissatisfied with the market equilibrium price. When that happens, they lobby government for price controls. This government intervention in the markets, designed to resolve a perceived problem in the market price, will create a different problem and is usually not advised by most economists.

## References and Further Readings

Congressional Budget Office (CBO). (2009, April). *The impact of ethanol use on food prices and greenhouse-gas emissions.* Retrieved July 6, 2009, from http://www.cbo.gov/ftpdocs/100xx/doc10057/04-08-Ethanol.pdf

Congressional Budget Office (CBO). (2009, November 23). *The costs of reducing greenhouse-gas emissions.* Retrieved November 30, 2009, from http://www.cbo.gov/ftpdocs/104xx/doc10458/11–23-GHG_Emissions_Brief.pdf

Energy Information Agency (EIA). (2009, September). *Quarterly coal report April–June, 2009.* Retrieved November 30, 2009, from http://www.eia.doe.gov/cneaf/coal/quarterly/qcr.pdf

Energy Information Agency (EIA). (n.d.). *Oil market basics: A primer on oil markets combined with hotlinks to oil price and volume data available on the Internet*. Available at http://www.eia.doe.gov

Federal Trade Commission. (2005). *Gasoline price changes: The dynamic of supply, demand, and competition.* Retrieved July 6, 2009, from http://www.ftc.gov/reports/gasprices05/050705gaspricesrpt.pdf

# 8

# Consumer Behavior

Frederick G. Tiffany
*Wittenberg University*

The demand curve is one half of the familiar supply-and-demand graph that determines price and quantity exchanged in a perfectly competitive goods market. This curve shows the relationship between the price of a commodity and the quantity that buyers are willing to buy at each given price, holding all other factors constant. The shape of the curve, as well as how it shifts when various other factors do change, is explained by the theory of consumer behavior. The most important aspect of consumer theory is its conclusion that (in almost all circumstances) the demand curve for a commodity is downward sloping. This relationship is the familiar "law of demand" first articulated directly by Alfred Marshall (1842–1924) in his *Principles of Economics* (Marshall, 1930, p. 99).

Consumer theory applies equally well to tangible goods and intangible services; the words *goods* and *commodities* are often used interchangeably. The theory does not, however, apply to the demand for factors of production, such as labor, which depends on the profit-maximizing behavior of firms.

Early consumer theory grew out of the work of several economists who developed the idea that the happiness or "utility" that a consumer receives from successive increments of a commodity declines. This idea has come to be known as the principle of "diminishing marginal utility." When consumers are deciding how to allocate a fixed budget among various commodities, diminishing marginal utility guides them to spend more on a commodity if its price falls relative to the prices of all other commodities. An argument for a downward-sloping demand curve can be made based on this principle. However, without a reliable way to measure utility, the law of demand rests on a shaky foundation. Fortunately, later economists developed a theory of the consumer from which the law of demand can be derived without reference to measurable utility.

Modern consumer theory is a mathematical theory of rational choice. Individuals or households are assumed to have the ability to rank all possible combinations or "bundles" of commodities they might possibly consume. They choose the highest ranked bundle they can afford given their income and the prices of the commodities. The impact of an increase in the price of a commodity on the quantity demanded by an individual buyer depends on that buyer's reaction to two separate but related effects: the increase in the price of the good relative to the prices of all other goods (the "substitution effect") and the decrease in the purchasing power of the consumer's budget (the "income effect"). The law of demand depends on these two effects, which can be expressed without reference to measurable utility.

This chapter explains the development of consumer theory from the introduction of the idea of utility through the modern theory. The following section traces the history of the argument that diminishing marginal utility is the basis for a downward-sloping demand curve. Subsequent sections provide a description of modern consumer theory based on a rank ordering of consumption bundles. The chapter concludes with references to areas of economics to which the consumer theory model is related.

## Diminishing Marginal Utility

Consumer behavior rests on the desires that buyers have for commodities. The earliest method for representing these desires came from the idea set forth by Jeremy Bentham (1748–1832) that every human action increases or diminishes the happiness of the one taking the action (Bentham, 1907, p. 2). Eating an apple may cause a person's happiness to increase, while mopping the kitchen

floor may cause that person's happiness to decrease. The increase or decrease in happiness from each of these actions is called the "utility" from taking that action. The desire that a person has for a commodity, therefore, is related to the utility that the consumer receives from consuming it. Bentham also introduced the idea of diminishing marginal utility when he suggested that "the quantity of happiness produced by a particle of wealth (each particle being of the same magnitude) will be less at every particle: the second will produce less than the first, the third less than the second, and so on" (Bentham, 1952, p. 113). This concept has come to be known as the diminishing marginal utility of wealth (or income). The idea is that as one spends wealth, the utility received for each additional dollar is not constant but declines. Note that Bentham did not refer directly to the marginal utility of consuming a particular good or service itself.

Jules Dupuit (1804–1866) did consider the utility of consuming a particular good. In his paper "On the Measurement of Utility of Public Works" (Dupuit, 1969), he argued that the utility of water purchased from a public water system declined as successive units were consumed. At a given price, people will only buy water for which the utility they receive exceeds that price. If the price falls, buyers will purchase additional water for uses that have a utility less than the higher price but more than the lower price (pp. 258–259). Because a lower price is necessary to induce the purchase of more water, the utility of successive units must be diminishing. Following Jean Baptiste Say (1767–1832), Dupuit measured utility implicitly as the amount consumers were willing to pay to purchase a good. He was quite clear on this point: "Hence the saying which we shall often repeat because it is often forgotten: the only real utility is that which people are willing to pay for" (p. 262). Dupuit illustrated his ideas by constructing a downward-sloping "curve of consumption," with price on the horizontal axis and quantity of water on the vertical (pp. 280–281). This curve is a forerunner of the modern demand curve.

It should be noted that Dupuit's (1969) argument did not rely on the direct measurement of the amount of pleasure or pain attributable to the water consumed. He measured utility indirectly as the amount of money it would be worth to a consumer. This concept is essentially what modern economists call the consumer's "willingness to pay."

Dupuit (1969), however, focused on only a single commodity. He did not explain how a consumer would allocate a fixed budget among several commodities. Between 1854 and 1874, four different economists independently made arguments that are equivalent to the now-familiar statement that consumers will maximize utility when they choose to divide their income among commodities in such a way that the marginal value of the last dollar (or any monetary unit) spent is equal for each one. The first of these was Herman Heinrich Gossen (1810–1858), whose work was published in 1854. He stated the proposition that "man obtains the maximum of life pleasure if he allocates all his earned money between the various pleasures . . . in such a manner that the last atom of money spent for each pleasure offers the same amount [intensity] of pleasure" (Gossen, 1854/1983, pp. 108–109). This statement is equivalent to the now-familiar mathematical rule that for any set of $n$ commodities, the budget of a utility-maximizing consumer should be spent such that

$$\frac{MU_i}{P_i} = \frac{MU_2}{P_2} = \cdots = \frac{MU_n}{P_n}. \tag{1}$$

In this equation, $MU_i$ and $P_i$ denote the marginal utility of Good $i$ and the price of Good $i$, respectively, and their quotient is the marginal utility of the last dollar spent on Good $i$. Assuming diminishing marginal utility and constant prices, if for any two Goods $i$ and $j$,

$$\frac{MU_i}{P_i} > \frac{MU_j}{P_j}, \tag{2}$$

then purchasing one dollar's worth less of Good $j$ will cause a loss of utility that is less than the utility gained from spending that dollar on Good $i$ instead. Therefore, the utility from consuming these goods is not maximized. It will be maximized only if there is no opportunity to change expenditures in this way. It should be noted that with only two commodities, the expression in Equation 1 can be written as

$$\frac{MU_1}{MU_2} = \frac{P_1}{P_2}. \tag{3}$$

The ratio of the marginal utilities is equal to the ratio of the prices.

William Stanley Jevons (1835–1882), Carl Menger (1840–1921), and Leon Walras (1834–1910) each presented an argument that was equivalent to the one made by Gossen, although in different forms. Menger's (1871/1950) work in *Principles of Economics* was the least overtly mathematical, and Walras's (1977) *Elements of Pure Economics* was the most. While Gossen's work came more than 15 years prior to the others, they did not seem to have been aware of it. In the preface to the second edition of Jevons's (1871/1965) book, *The Theory of Political Economy,* he describes his surprise at discovering the existence of Gossen's work and gives Gossen credit for having "completely anticipated" his own work (p. xxxv).

Although neither Gossen nor those who followed him made it, the familiar textbook argument for a downward-sloping demand curve based on diminishing marginal utility is a direct implication of Equation 1. Suppose that the price of Good 1 falls, holding everything else in the equation constant. Then the marginal utility of the last dollar spent on Good 1 would exceed the marginal utilities of the last dollar spent on all other goods. The consumer, therefore, would increase the amount spent on Good 1 and decrease the amount spent on all other goods. If diminishing marginal utility applies to all of the goods, purchasing

more units of Good 1 would lower its marginal utility, while purchasing less of the other goods would raise their marginal utilities. This process would continue until the equation was once again satisfied. Thus, because of diminishing marginal utility, the consumer would respond to a decrease in the price of Good 1 by demanding more of it. The demand curve for Good 1 would be downward sloping. The same reasoning can be used to show that an increase in the price of Good 1 would lead the consumer to desire less of it.

Alfred Marshall's (1930) *Principles of Economics* is one of the classics of the field and was a standard textbook for many years. In his chapter on "Wants in Relation to Activities," Marshall used the idea of diminishing marginal utility as the basis for a demand curve. His argument was similar to that of Dupuit (1969). Because of diminishing marginal utility, Marshall stated,

> The larger the amount of a thing that a person has the less, other things being equal, (*i.e.* the purchasing power of money, and the amount of money at his command being equal) will be the price which he will pay for a little more of it: or in other words, his marginal demand price for it diminishes. (p. 95)

Marshall (1930, pp. 96–99) went on to derive a demand schedule and a demand curve for both the individual and the market. He also presented a version of the argument that utility is maximized when the marginal utility of the last monetary unit spent on each good is equal (p. 118).

It is important to note, however, that Marshall (1930) stated clearly that "desires cannot be measured directly, but only indirectly by the outward phenomena to which they give rise . . . in the price which a person is willing to pay" (p. 92). In a footnote, he emphasized the point, saying, "It cannot be too much insisted that to measure directly . . . either desires or the satisfaction which results from their fulfillment is impossible, if not inconceivable" (p. 92). Thus, the way was opened for a theory that does not rely on the concept of measurable utility at all.

John Hicks (1904–1989) laid the groundwork for modern consumer theory in *Value and Capital* (Hicks, 1946). He took the idea of an indifference map, developed by Vilfredo Pareto (1848–1923) in his *Manual of Political Economy* (Pareto, 1909/1971, pp. 118–122), and used it to show that a consumer can choose the optimal combination of commodities without reference to measurable utility at all. (Pareto, in turn, credited the first construction of indifference curves to F. Y. Edgeworth's [1881/1967] *Mathematical Psychics,* pp. 21–22.) For Hicks, a utility number "only tells us that the individual prefers one particular collection of goods to another particular collection; it does not tell us *by how much* the first collection is preferred to the second" (p. 17). Paul Samuelson (1915–2009), who provided the mathematical foundation for modern consumer theory, concluded that "it is only necessary that there exist an *ordinal* preference field" to derive demand curves (Samuelson, 1979, p. 93). The concept of diminishing marginal utility had run its course.

## Modern Consumer Theory

Modern consumer theory presents an argument for a downward-sloping demand curve that rests not on diminishing marginal utility but rather on what have become known as the income and substitution effects. The theory depends on a minimal number of assumptions about a consumer's ability to rank various combinations or "bundles" of commodities and some basic characteristics of all consumers' rankings. This section will set forth this argument in detail.

### Assumptions Regarding Preferences

The basic building blocks of the theory are *consumption bundles.* Suppose that there are only two goods in an economy, $x$ and $y$. A consumption bundle is an ordered pair of numbers $(x, y)$ in which each number represents an amount of the respective goods. Thus, (2, 1) represents two units of $x$ and one of $y$. A consumption bundle can include any number of goods, but using more than two introduces mathematical complications that are unnecessary for developing the basic theory. A bundle such as $(x_1, y_1)$ can be represented as a point in the commodity space defined by a set of $x$, $y$ axes. The set of all bundles in the positive quadrant defined by these axes is called a *commodity space.* It is assumed that commodities are infinitely divisible and that the space extends infinitely to the northeast.

Consumers are assumed to be able to place bundles into a rank ordering, so that for any pair of bundles $A$ and $B$, they will be able to say one of three things: They prefer $A$ to $B$, they prefer $B$ to $A$, or they are indifferent between the two. In order for these comparisons to be logically consistent, three things must be true:

1. Consumer preferences must be *complete.* Consumers must be able to state a relationship between any two bundles they are presented with. They may say that they prefer $A$ to $B$, that they prefer $B$ to $A$, or that they are indifferent between the two. They cannot say that they do not know what their preference is regarding the two. This assumption simply amounts to saying that consumers know their own preferences.
2. Consumer preferences must be *reflexive.* Consumers must be able to recognize when two bundles are identical and be indifferent between them.
3. Consumer preferences must be *transitive.* If a consumer prefers $A$ to $B$ and $B$ to $C$, then the consumer must also prefer $A$ to $C$. If the consumer is indifferent between $A$ and $B$, as well as between $B$ and $C$, then the consumer must be indifferent between $A$ and $C$. The same must be true for combinations of these relationships: For example, if the consumer prefers $A$ to $B$ and is indifferent between $B$ and $C$, then $A$ must be preferred to $C$.

Based on these three assumptions, the bundles in the commodity space can be organized into *indifference curves.* An indifference curve is a set of all bundles among which a consumer is indifferent. The set of all indifference curves for a consumer is called an *indifference map.* Each consumer has a unique indifference map that will determine the consumer's demand for each of the goods in the commodity space.

While these assumptions are all that are necessary for deriving an indifference map, two more are required to establish its familiar shape. These assumptions are not about logical consistency but about how consumers view their consumption decision. Continuing the previous numbering, they are as follows:

4. Consumers *prefer more to less.* If bundle $A$ has no less of each good than bundle $B$ and more of at least one good, then the consumer prefers $A$ to $B$. This assumption is often referred to as "monotonicity" of preferences.
5. Consumers have a *diminishing marginal rate of substitution* (MRS) between any two goods. The amount of one good ($y$) that a consumer is willing to give up to attain an additional unit of another good ($x$) gets smaller as the amount of $x$ gets larger.

## Graphical Representation of Indifference Maps

Taken together, the last two assumptions allow the drawing of an indifference map, represented in Figure 8.1a by three of the consumer's indifference curves, $I_1$, $I_2$, and $I_3$. The curves are downward sloping, with the slope becoming less steep as the amount of $x$ increases. Such curves are said to be convex to the origin. The slope along an indifference curve $\frac{\Delta y}{\Delta x}$ is the MRS because it shows the amount of $y$ that a consumer will give up ($\Delta y$) in exchange for an increase in $x$ ($\Delta x$) while staying on the same indifference curve. The indifference curves must be downward sloping because of the assumption that more is preferred to less. An upward-sloping indifference curve would violate this assumption because for every bundle $(x_1, y_1)$ on the curve, there would be another bundle $(x_2, y_2)$ in which $x_2 > x_1$ and $y_2 > y_1$. But the assumption that more is preferred to less implies that $(x_2, y_2)$ would be preferred to $(x_1, y_1)$, which violates the definition of an indifference curve.

The assumptions of transitivity and monotonicity also imply that no two indifference curves may cross one another. Suppose that two separate indifference curves $I_1$ and $I_2$ do cross. Then, because these curves are distinct, there must be at least one bundle $A$ on $I_1$ that is preferred to a

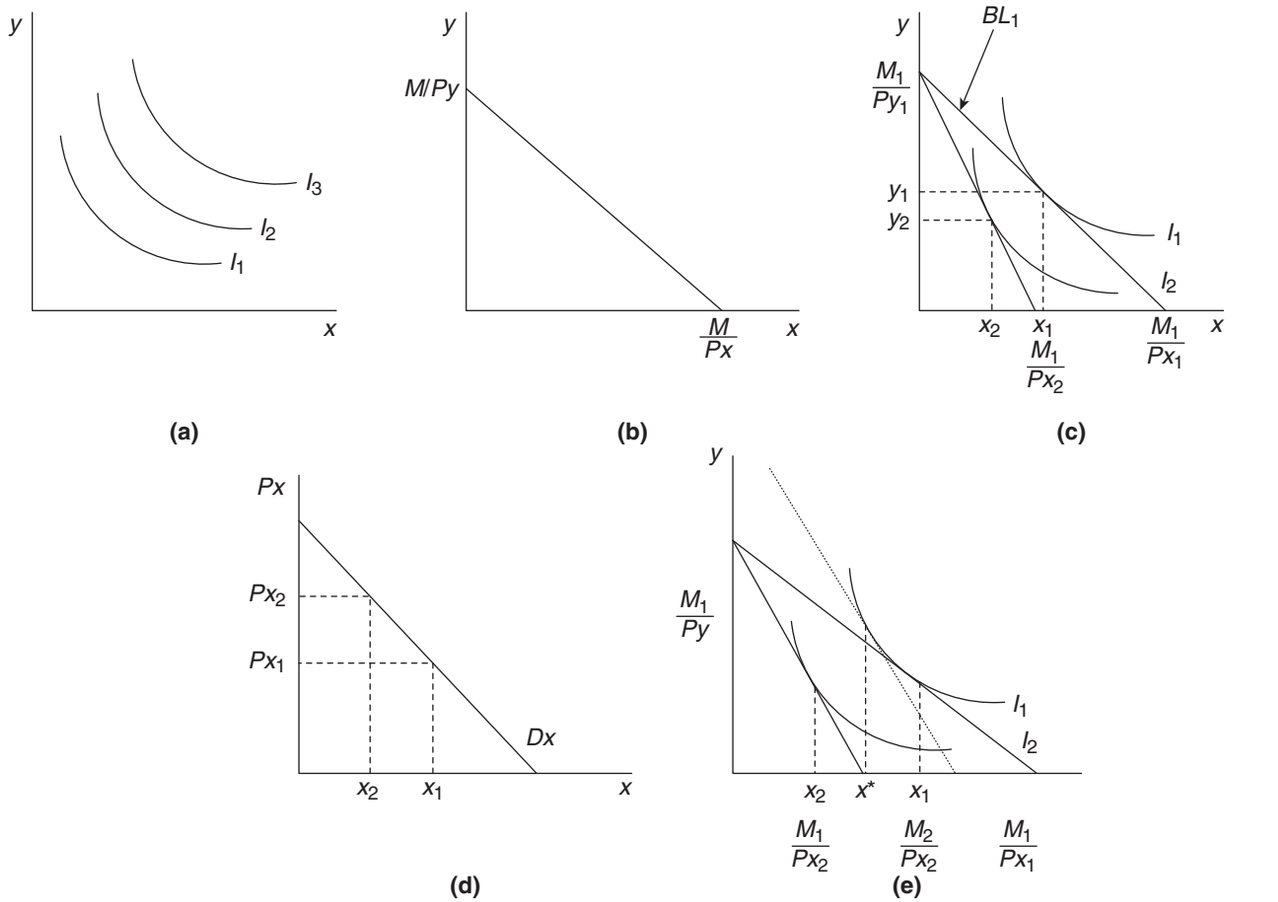


**Figure 8.1** Consumer Behavior

bundle $B$ on $I_2$. However, there must also be a bundle $C$, where the curves cross, which is on both curves. The consumer must be indifferent between $A$ and $C$ and between $B$ and $C$. However, by transitivity, if the consumer is indifferent between $A$ and $C$ and between $B$ and $C$, then the consumer must be indifferent between $A$ and $B$. But because $A$ is preferred to $B$, there is a logical contradiction because the consumer cannot prefer $A$ to $B$ and also be indifferent between these two bundles.

## Utility Functions

The concept of utility played no role in the derivation of indifference curves in the previous sections. Indifference curves depend only on consumers' ability to rank bundles consistently. However, it is convenient to give a numerical scale to the ranking of bundles. Such a scale is called a *utility function.* More precisely, a utility function assigns a number to each possible consumption bundle such that if $A$ is preferred to $B$, then $A$ is given a higher number. If the consumer is indifferent between $A$ and $B$, the utility function gives them the same number. Any function that assigns numbers in this way is acceptable. For example, suppose that bundles $A$ and $B$ are assigned utility numbers of 2 and 4, respectively. $B$ is preferred to $A$. If a constant, say 2, is added to each number, $B$ is still preferred to $A$. The same can be said for multiplying each number by a constant, taking each to a power or some combination of these "monotonic transformations." Consequently, an infinite number of utility functions can represent a given set of preferences.

A utility function can be applied to an indifference map by assigning a number to each indifference curve. Any set of numbers will do as long as curves farther from the origin receive higher numbers. The reason is as follows. Suppose that indifference curve $I_2$ is farther from the origin than $I_1$. Then, there is at least one bundle $B$ on $I_2$ that has more of both goods than some bundle $A$ on $I_1$. By monotonicity, $B$ is preferred to $A$ and receives a higher utility number. By the definition of an indifference curve, all bundles on $I_2$ have the same number as $B$, and all bundles on $I_1$ have the same (lower) number as $A$.

Once a utility function is established, it is sometimes convenient to use the idea of marginal utility within the numerical scale provided by the function. In this context, marginal utility is the change in the utility number when the amount of one of the goods is increased by a small amount, holding the amount of the other good constant (i.e., for a move horizontally or vertically from one indifference curve to another). Marginal utility can be increasing or diminishing for the same set of indifference curves depending on the particular utility function that is chosen. If a horizontal line is drawn that crosses successive indifference curves at intervals of one unit of $x$, the utility numbers where the line crosses three successive indifference curves could be 3, 5, 6 or they could be 3, 7, 12. In the first case, marginal utility would be decreasing, and in the second, it would be increasing. But the underlying preferences represented by the indifference curves would be the same. Thus, diminishing marginal utility is not an intrinsic property of preferences that meet the basic assumptions. Rank order, not the intensity of satisfaction, is all that matters.

There is, however, an important relationship between the marginal rate of substitution and the utility function that represents a particular set of preferences. Consider a movement along an indifference curve from bundle $(x_1, y_1)$ to bundle $(x_2, y_2)$. Suppose that this move takes place in two stages. First, $x$ increases by a small amount $\Delta x$, holding $y$ constant. Because more is preferred to less, the utility number for the new bundle $(x_1 + \Delta x, y_1)$ must be larger than it was for $(x_1, y_1)$. (Note that $x_2 = x_1 + \Delta x$.) The change in utility can be represented as $MUx \bullet \Delta x$, which is the marginal utility for a small change in $x$ multiplied by that change in $x$. A return to the original indifference curve requires $y$ to decrease by a small amount of $\Delta y$, holding $x$ constant at $x_2$. The change in utility from this move is $MUy \bullet \Delta y$. The sum of these two changes in utility must be zero. Thus,

$$MUx \bullet \Delta x + MUy \bullet \Delta y = 0. \tag{4}$$

This equation is algebraically equivalent to

$$\frac{\Delta y}{\Delta x} = -\frac{MU_x}{MU_y}. \tag{5}$$

The left-hand side of Equation 5 is the marginal rate of substitution along the indifference curve. Therefore, the MRS can be expressed as the ratio of the marginal utilities of $x$ and $y$. This relationship is true for any utility function that accurately represents the consumer's preferences and will be important in connecting modern consumer theory with its predecessor.

## Budget Lines

The goal of the consumer is to choose the affordable commodity bundle with the largest utility number, given the consumer's income and the prices of the goods. Therefore, it is necessary to characterize the consumer's budget constraint. The idea is simple: The sum of the amounts spent on each good must not exceed the consumer's income. The commodity bundles that satisfy this inequality are called the *budget set.* However, because consumers prefer more to less, they will never purchase a bundle that does not use up all of their income. Therefore, the only bundles that might be chosen are those on the *budget line,* which can be expressed algebraically as

$$Px \bullet x + Py \bullet y = M, \tag{6}$$

where $Px$ and $Py$ are the prices of the goods and $M$ is the consumer's income. The budget line can be graphed by solving for $y$ in terms of $x$ to get

$$y = \frac{M}{P_y} - \left(\frac{P_x}{P_y}\right) \bullet x. \tag{7}$$

The budget line can be graphed as in Figure 8.1b, with the endpoints $\frac{M}{py}$ and $\frac{M}{px}$ representing the amounts of $y$ and $x$, respectively, that could be purchased if a consumer purchased only one good. The slope of the budget line is the opportunity cost of one more unit of $x$ in terms of $y$. If, for example, $Px = 3$ and $Py = 1$, then $x$ is three times as expensive as $y$, so a consumer must sacrifice three units of $y$ to purchase one more unit of $x$.

## Choosing the Best Consumption Bundle

For a given set of prices, income, and indifference curves, a consumer will choose the bundle that is on the indifference curve farthest from the origin but still touching the budget line. Suppose that the prices of $x$ and $y$ are $Px_1$ and $Py_1$, respectively, and the consumer's income is $M_1$. Then the consumer's budget line can be represented as $BL_1$ in Figure 8.1c. Given this budget line and the consumer's indifference map, the bundle $(x_1, y_1)$ on indifference curve $I_1$ is the best the consumer can do. Note that as long as this highest indifference curve has no flat spots, no sharp corners, and does not touch the axes, it will be *tangent* to the budget line at $(x_1, y_1)$, which means that the budget line and indifference curve have the same slope. As shown above, the slope of the indifference curve can be expressed as the ratio of the marginal utilities of $x$ and $y$ (5), while the slope of the budget line is the ratio of the prices (7). Setting these equal gives

$$\frac{MU_x}{MU_y} = \frac{Px}{Py}. \tag{8}$$

As shown in Equation 3 above, this is an implication of Gossen's rule. With only two goods, it is equivalent to the consumer allocating income among goods so that the marginal utility of the last dollar spent on each good is the same, shown in Equation 1 above. Thus, modern consumer theory comes to the same conclusion that Gossen and others reached. The difference is that the utility numbers have no intrinsic meaning. They only show a rank ordering of the bundles. As shown above, if marginal utility is diminishing, a downward-sloping demand curve can be derived from these conditions. However, the same indifference curves could be ranked by a utility function that has increasing marginal utility. The next section shows how a demand curve can be derived without any reference to specific utility numbers.

## Demand Curves

A demand curve shows the quantity that a consumer will demand at each given price, with the prices of other goods, income, and preferences held constant. In Figure 8.1c, $Px_1$ and $x_1$ meet this requirement because $x_1$ is the quantity that the consumer demands at $Px_1$, given the price of the other good $Py_1$, income $M$, and the preferences captured by the indifference map. So $(Px_1, x_1)$ is a point on the demand curve for $x$, shown as $Dx$ in Figure 8.1d. To derive a complete demand curve, the price of $x$ needs to be varied, holding all other aspects of the graph constant. Suppose that the price of $x$ increases to $Px_2$. Then, as shown in Figure 8.1c, the budget line rotates inward. The vertical intercept does not change, but the horizontal intercept, the maximum amount of $x$ the consumer could buy, becomes smaller. The slope of the budget line becomes steeper because $Px$ increases relative to $Py$, which remains constant, as do income and the indifference curves. As shown in the figure, the consumer would now choose $x_2 < x_1$. $(Px_2, x_2)$ is a second point on the demand curve for $x$. The entire demand curve can be derived by simply changing $Px$, rotating the budget line, and finding the new maximizing amount of $x$ for each price.

## Engel Curves and Cross-Price Demand Curves

The basic theory of demand asserts that changes in income, preferences, or the price of other goods cause a demand curve to shift left or right. The theory of the consumer can show the effects of changes in income or the price of the other good using *Engel curves* and *cross-price demand curves.* An Engel curve shows the amount of a good that is demanded at different levels of income, holding preferences and the prices of both goods constant. It can be derived from Figure 8.1c by changing income, holding the prices of both goods constant. (This derivation is not shown in the figure.) For each different level of income, the budget line moves parallel left or right, and a new best bundle is obtained. For each budget line, the amount of $x$ demanded can then be paired with the income for that budget line and these pairs plotted with $M$ on one axis and $x$ on the other. An Engel curve for a *normal good* is upward sloping because demand increases as income increases, while for an *inferior good,* where demand decreases as income increases, the curve is downward sloping. Rice is a common example of an inferior good. If a consumer has a low income, the consumer probably eats lots of rice and little meat. If rice is viewed as "inferior" to meat, then an increase in the consumer's income, ceteris paribus, is likely to cause the consumer to demand more meat and less rice.

A cross-price demand curve shows the amount of one good demanded at various prices of the other good, holding the price of the first good and income constant. It can be derived for good $y$ from Figure 8.1c by pairing the amount of $y$ demanded from each bundle with the price of $x$, as the price of $x$ changes. Thus, the pairs $(Px_1, y_1)$ and $(Px_2, y_2)$ would be on this curve (not shown). A cross-price demand curve is upward sloping if the two goods are *substitutes* (as the price of $x$ increases, the consumer demands more $y$ to substitute for $x$) and downward sloping if they are *complements* (as the price of $x$ rises, the consumer wants less $y$ to consume with less $x$). For many people,

hamburgers and hot dogs are substitutes, while peanut butter and jelly are complements.

## Income and Substitution Effects

The downward slope of a demand curve depends on how the consumer reacts to two aspects of a change in the price of a good. The first is the change in the price of the good relative to the price of the other good. This change is represented by the change in the slope of the budget line, and the consumer's reaction to it is called the *substitution effect.* The other aspect is the change in the purchasing power of the consumer's income. This is roughly represented by the change in the size of the budget set, and the consumer's reaction to it is called the *income effect.* Showing these two effects requires a careful analysis of the consumer's move from $x_1$ to $x_2$ when the price changes.

This analysis is shown in Figure 8.1e. The best bundles at prices $Px_1$ and $Px_2$ are $(x_1, y_1)$ and $(x_2, y_2)$ as before. However, the movement between them is now broken into two parts, one due to the substitution effect and the other to the income effect. To separate these two, the following thought experiment is necessary. Suppose that in response to the increase of the price of $x$, the consumer were given just enough additional income to shift the budget line so that it would again be tangent to the original indifference curve, $I_1$. This increase from $M_1$ to $M_2$ would compensate for the loss of purchasing power caused by the increase in the price of $x$. However, the increase in the price of $x$ relative to the price of $y$ means that the slope of this augmented budget line is steeper than the slope of the original one. Consequently, the consumer would choose to consume $x^*$ rather than $x_1$ as a result of the change in relative prices. This bundle contains less $x$ and more $y$ than the original bundle. The consumer, therefore, would substitute some $y$ for $x$ simply because of the change in relative prices. This change in the amount of $x$ demanded $(x_1 - x^*)$ is called the *substitution effect.* It is always negative (i.e., the consumer always substitutes away from the good that has become relatively more expensive). But the increase in income was only a thought experiment. In fact, the consumer's real income has fallen. The effect on the amount of $x$ demanded caused only by the drop in real income is measured by finding the difference between $x^*$ and $x_2$. This difference is called the *income effect.* If the good is a normal good, then a decrease in income will cause the consumer to choose less $x$. Because both the income and substitution effects result in less $x$ being demanded, the demand curve is downward sloping.

However, the good could be an inferior good. In that case (not shown), the fall in real income, taken alone, would lead the consumer to demand more $x$. The income and substitution effects would then be working in opposite directions, and $x_2$ would be to the right of $x^*$. But as long as $x_2 < x_1$, the combined effect would still be for the quantity demanded of $x$ to fall. The demand curve would again be downward sloping but steeper than it would be for a normal good.

Finally, it is possible that the good could be so inferior that the increase from $x^*$ to $x_2$ caused by the lower real income would be greater than the decrease from $x_1$ to $x^*$ caused by the change in relative prices. In that case, the demand curve would be upward sloping, with $x_2 > x_1$ (not shown). The law of demand would not hold. A good with this property is called a Giffen good, named after Robert Giffen (1837–1910). While a Giffen good is a theoretical possibility, it is hard to imagine a real-world example. In the example above, rice would have to be so inferior that faced with the loss in purchasing power caused by an increase in the price of rice, the consumer would actually reduce meat consumption to consume more rice. Although indifference curves can be drawn to represent such preferences, they seem highly unlikely in the real world. But the important point is that demand curves are generally considered to be downward sloping precisely because such preferences are not likely.

The law of demand depends on the nature of the preferences that shape the indifference curves, not on the numbers assigned to them. The demand curve for a good is downward sloping if the good is normal or, in the case of an inferior good, if the income effect is less than the substitution effect. Diminishing marginal utility plays no part at all.

## Revealed Preference

An interesting twist to consumer theory was proposed by Paul Samuelson (Samuelson, 1938). He suggested that consumer preferences could be uncovered by observing what consumers actually choose from among bundles they can afford. Suppose that there are two consumption bundles, $A$ and $B$, on the consumer's budget line. If the consumer is observed to choose $A$, then it is "revealed" that the consumer prefers $A$ to $B$ because $A$ was chosen when both were affordable. Such a comparison can be made only between bundles that are both in the budget set. However, by changing the budget line (and set), comparisons can be made among enough different bundles that indifference curves can be sketched out. It is important to note, however, that this approach does not represent a different consumer theory. The assumptions given above are all assumed to hold. What is different is the lens through which the data of preferences are perceived.

## Intertemporal Choice

It is straightforward to apply consumer theory to consumption in different periods of time. The commodity bundles can simply be reinterpreted as including the amount of consumption of a standard good in different time periods. Thus, in Figure 8.1a, consumption in the first period (e.g., this year) would be measured along the

horizontal axis, while consumption in the second period (next year) would be measured along the vertical axis. The slope of the budget line, which represents the price of a unit of consumption this year in terms of forgone consumption next year, would be the real interest rate. By consuming a dollar's worth of the commodity now, the consumer is giving up the additional amount that could be purchased with the interest earned by saving until next year. This model can be extended mathematically to include any number of time periods.

### Choice When Future Consumption Is Uncertain

Households may need to make choices when the outcome of those choices is not certain. The most common situations involve gambles, particularly the gamble that one takes when purchasing a financial asset. Rather than being able to know for certain that an asset will allow a given amount of consumption in the future, the household may calculate the probabilistic expected value of the amount of consumption that the asset will yield. A household's attitude toward risk will determine how it views these gambles. Households may be "risk averse," meaning they would prefer a certain outcome to a gamble with the same expected value as the certain outcome. They may be "risk loving," meaning they prefer the expected value of the gamble to the certain outcome, or they may be "risk neutral," meaning they are indifferent between the two. Their attitudes toward risk will, in turn, affect the premium that must be paid to accept risk or that they are willing to pay to avoid risk. These considerations are important in Chapter 20 ("Asset Pricing Models") and Chapter 21 ("Portfolio Theory and Investment Management") in this volume.

### Welfare Economics

The effect of economic decisions on the well-being of consumers and producers is addressed in the field of economics referred to as welfare economics. Models in this field often make use of the idea of a "compensated demand curve." This curve relates price and quantity demanded when only the substitution effect is taken into account. A precise measurement of consumer's surplus, the amount by which a consumer's willingness to pay exceeds the amount the consumer pays for a unit of a good, requires using a compensated demand curve.

### Labor Markets

Even though consumer theory is concerned with commodity markets, the model presented above is applied to the supply of labor in Chapter 14, "Labor Markets" (this volume). Even though consumer theory is concerned with commodity markets, the model presented above is applied to the analysis of labor supply decisions as well. In the standard model of labor supply, households are assumed to divide their time between hours spent working for income and hours "consumed" as leisure (not earning income). The model employs indifference curves for bundles consisting of a number of hours of leisure and an amount of income earned from labor. Households choose the combination of leisure and income on the highest indifference curve touching their budget constraint. The constraint is composed of bundles of income and leisure hours that can be obtained under a given wage rate. The total number of possible hours in a day minus the hours of leisure chosen equals the hours of labor supplied.

### General Equilibrium

Models that include many markets at once are used to determine the conditions under which there exists a set of prices that will clear all of the markets simultaneously. Demand in goods markets is generally modeled using the tools of indifference curves and budget constraints described in this chapter. This volume contains a chapter on "Supply, Demand, and Equilibrium" (Chapter 7).

### Challenges to Traditional Theory

The model of consumer behavior presented in this chapter is based on strictly rational behavior. In recent decades, the new fields of behavioral economics and experimental economics have challenged traditional models by suggesting that there are bounds to the rationality that consumers can exercise. These related fields look for explanations of data that do not fit the standard theory presented in this chapter. There are separate chapters on each of them (Chapters 84 and 85) in this volume.

## Conclusion

The law of demand is fundamental to economic models of markets. The theory of consumer behavior developed in the nineteenth and early twentieth centuries derived a downward-sloping demand curve based on the concept of measurable diminishing marginal utility. Beginning with Hicks, however, modern consumer theory has shown that the law of demand can be derived based only on the consumer's ability to rank-order alternative consumption bundles and choose the most preferred bundle that falls within the set of bundles the consumer can afford given income and the prices of the goods. Thus, the demand curve familiar to all economics students has a rigorous basis in theory. The model is useful in many areas of economics, but its reliance on strictly rational behavior recently has been called into question and will likely be the subject of much future research.

## References and Further Readings


Bentham, J. (1907). *An introduction to the principles of morals and legislation.* New York: Oxford University Press. (Original work published 1789)

Bentham, J. (1952). *Jeremy Bentham's economic writings* (Vol. 1). London: Royal Economic Society.

Brue, S. (2000). *The evolution of economic thought* (6th ed.). Mason, OH: Thompson South-Western.

Dupuit, J. (1969). On the measurement of the utility of public works. In K. Arrow & T. Scitovsky (Eds.), *Readings in welfare economics* (pp. 255–283). Homewood, IL: Irwin.

Edgeworth, F. Y. (1967). *Mathematical psychics.* New York: Augustus M. Kelley. (Original work published 1881)

Ekelund, R., & Hebert, R. (1990). *A history of economic theory and method* (3rd ed.). New York: McGraw-Hill.

Gossen, H. H. (1983). *The laws of human relations and the rules of human action derived therefrom.* Cambridge: MIT Press. (Original work published 1854)

Hicks, J. R. (1946). *Value and capital* (2nd ed.). Oxford, UK: Oxford University Press.

Jevons, W. S. (1965). *The theory of political economy.* New York: Augustus M. Kelley. (Original work published 1871)

Landreth, H., & Colander, D. (2002). *History of economic thought* (4th ed.). Boston: Houghton Mifflin.

Marshall, A. (1930). *Principles of economics.* London: Macmillan.

Menger, C. (1950). *Principles of economics.* Glencoe, IL: Free Press. (Original work published 1871)

Nicholson, W. (2002). *Microeconomic theory* (8th ed.). Mason, OH: Thomson South-Western.

Pareto, V. (1971). *Manual of political economy.* New York: Augustus M. Kelley. (Original work published 1909)

Pindyck, R., & Rubinfeld, D. (2009). *Microeconomics* (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Samuelson, P. (1938). A note on the pure theory of consumers' behavior. *Economica, 5,* 61–71.

Samuelson, P. (1979). *Foundations of economic analysis.* New York: Atheneum.

Walras, L. (1977). *Elements of pure economics.* Fairfield, CT: Augustus M. Kelley.

# 9

# Demand Elasticities

Ann Harper Fender
*Gettysburg College*

Economic lore depicts Alfred Marshall leaping in 1881 from the low roof of the Oliva Hotel while on vacation in Palermo, Italy; Marshall, the legend continues, ran through the town's streets shouting, "Eureka, I've found it!" The legend has his excitement stemming from his discovery of a simple formalization for the concept of elasticity (Keynes, 1963).[1] Marshall was not the first to incorporate something like elasticity in his economic analysis. He was familiar with the work of mathematical economists Augustin Cournot and Johann von Thünen, both of whom developed theories of firm behavior earlier in the nineteenth century and arguably hinted at elasticity. The classical economist John Stuart Mill also discussed the impact of changes in price on quantity consumed and on the impact of tariffs on output and prices; Fleming Jenkin had alluded to the elasticity concept in 1870 (Ekelund & Hebert, 1990).

Marshall moved beyond general representations of the response of quantity demanded to a change in price. Specifically, he defined price elasticity of demand as the percentage change in quantity demanded divided by percentage change in price. It was not until 1890, when the first edition of his *Principles of Economics* appeared, that Marshall published his work on elasticity along with the many other contributions to economics that he had developed over several decades. In addition to being precise, his definition of elasticity had the benefit of being independent of the units in which price and quantity are measured, as illustrated in the next section, which also illuminates the technical rules for calculating arc and point price elasticities of demand and relates these to the geometry of demand curves. The third section explores the relationship between firm revenues and price elasticities of demand. Market power is related to price elasticity of demand, as the fourth section considers, along with the power of firms to price discriminate. The fifth section considers other demand elasticities, the factors that influence them, and their significance. Marshall extended the elasticity concept to the demand for inputs, described in the sixth section. Both classical and neoclassical economists were concerned with moral and social issues; their grappling with the concept of elasticity often resulted from concerns with how strong an effect particular public policies would have. In the seventh section, we consider how elasticity of demand matters in economic policy. Although the concept of elasticity of demand was originated over a century ago, limits on computational ability prior to the computer age made it more valuable as a theoretical than as an empirical concept. Improved data collection and the advent of high-speed computers dramatically increased the numerical estimates of elasticity, as the concluding section indicates.

## Technical Rules for the Calculation of Elasticities

Microeconomic theory develops the demand for a product or service primarily as a function of its own price, the prices of related goods and services, and income:

$$Q_d = f(p, p_r, M),$$

where $p$ is own price, $p_r$ is price of related good(s), and $M$ is income.

The definition of an elasticity is the percentage change in the dependent variable divided by the percentage change in the independent or causal variable. Thus, the price elasticity of demand is calculated as percentage change in quantity demanded divided by percentage change in price, holding all other independent variables constant:

$$\varepsilon_p^d = \{(\text{change in quantity})/(\text{base quantity})\} \bullet 100 \div \{(\text{change in price})/(\text{base price})\} \bullet 100$$

$$= \frac{\Delta Q}{Q} / \frac{\Delta P}{P}$$

$$= \frac{\Delta Q}{\Delta P} \bullet \frac{P}{Q}.$$

This also equals $\frac{d(\log Q)}{d(\log P)}$.

## Arc Elasticity

This formula for elasticity between two distinct points on the demand curve raises the question as to what is the base $Q$ and base $P$: the price/quantity combination before or after the change? To avoid ambiguity and to give the same percentage changes whether price is rising or falling, the average of the prices and the average of the two quantities are frequently used. This yields the computational formula for price elasticity of demand:

$$\varepsilon_p^d = \frac{\Delta Q}{\Delta P} \bullet \frac{\frac{P_1+P_2}{2}}{\frac{Q_1+Q_2}{2}} = \frac{\Delta Q}{\Delta P} \bullet \frac{\sum P}{\sum Q}.$$

A (relatively) simple computational formula for the price elasticity when moving from one point on the demand curve to a second point on the demand curve is then the (change in quantity divided by change in price) multiplied by the sum of prices divided by sum of quantities. This yields an arc elasticity of demand, measured for a given movement along the curve.

Consider the demand schedule in Table 9.1.

**Table 9.1** Demand Schedule

| *Price* | *Quantity* |
|---|---|
| \$9 | 0 |
| \$8 | 1 |
| \$7 | 2 |
| \$6 | 3 |
| \$5 | 4 |
| \$4 | 5 |
| \$3 | 6 |
| \$2 | 7 |
| \$1 | 8 |
| \$0 | 9 |

The arc elasticity for a change in price from \$8, with $Q = 1$, to \$7 with $Q = 2$, is

$$\varepsilon_p^d = \left\{\frac{(2-1)}{(7-8)}\right\} \bullet \left\{\frac{(7+8)}{(2+1)}\right\} = -5$$

Were price expressed in cents rather than dollars, the elasticity coefficient would be

$$\varepsilon_p^d = \left\{\frac{(2-1)}{(700-800)}\right\} \bullet \left\{\frac{(700+800)}{(2+1)}\right\} = -5.$$

Whether price is measured in dollars or cents, the percentage change in price is the same because units cancel; the elasticity coefficient is independent of the units of measurement.

Economic theory deduces that the demand curve is (usually) downward sloping; therefore, the price elasticity of demand will be negative: A fall in price will lead to a rise in quantity. Often, rather than expressing this elasticity as a negative number, its absolute value is taken and the resulting positive number is considered the value of the price elasticity. If the quantity is highly responsive to a fall in price, the percentage increase in quantity will be large relative to the percentage drop in price. The numerator of the elasticity will be larger than the denominator (in absolute value). The resulting absolute value of the price elasticity of demand will exceed 1, and the demand is elastic along that price range. If the absolute value of the elasticity coefficient is not taken, the demand curve is elastic when the coefficient is less than −1 (e.g., −5 is less than −1).

When price falls from \$5 to \$4, the elasticity coefficient is

$$\varepsilon_p^d = \frac{5-4}{4-5} \bullet \frac{5+4}{4+5} = -1.$$

The price decrease is just offset by a proportional quantity increase, resulting in an elasticity coefficient of −1, or 1 if absolute value is taken. The demand is unitary elastic for this price change.

With a price decrease from \$4 to \$3, the price elasticity of demand is

$$\varepsilon_p^d = \frac{6-5}{3-4} \bullet \frac{3+4}{6+5} = -\frac{7}{11}.$$

For this price change, the percentage or proportional change in quantity demanded is small relative to the percentage change in price (in the opposite direction). The resulting elasticity coefficient is greater than –1 or, in absolute value, less than 1; demand is inelastic for this price change. Table 9.2 summarizes these elasticities.

## Point Elasticity

A graph (see Figure 9.1) of the demand schedule in Table 9.1 illustrates the relationship between price and amount consumers are willing to buy, holding all other

**Table 9.2** Own Price Elasticities of Demand

| *State of Demand* | *Price Elasticity Coefficient:* $\varepsilon$ | *Absolute Value of Price Elasticity Coefficient:* $\lvert\varepsilon\rvert$ |
|---|---|---|
| Price elastic | $\varepsilon < -1$ | $\varepsilon > 1$ |
| Unitary price elastic | $\varepsilon = -1$ | $\varepsilon = 1$ |
| Price inelastic | $\varepsilon > -1$ | $\varepsilon < 1$ |

relevant variables constant. By convention, economists follow the Marshallian practice of plotting quantity on the horizontal axis and price per unit on the vertical axis. This graphs the inverse demand curve

$$P = f(Q).$$

Rather than considering the price elasticity of demand along an arc of the demand curve, elasticity can be calculated at a single point, for example, point *A*. A point price elasticity of demand calculates the impact on quantity when the change in price becomes very small and hence equals the derivative of quantity with respect to price times the ratio of *P* to *Q* at that point:

$$\frac{\Delta Q}{\Delta P} \bullet \frac{\sum P}{\sum Q} = \left\{ \lim_{\Delta P \to 0} \frac{\Delta Q}{\Delta P} \right\} \bullet \frac{P}{Q} = \frac{dQ}{dP} \bullet \frac{P}{Q}.$$

The slope of the demand curve is $\frac{\Delta P}{\Delta Q}$. The first term in the price elasticity of demand is the inverse of this slope. For an arc elasticity, it is the inverse of the slope for a distinct movement along the curve; for a point elasticity, it is the inverse of the slope of a straight line tangent to the demand curve at that point. The price elasticity of demand can be written as

$$\varepsilon_p^d = \left\{ \frac{1}{(\text{slope of demand curve for given price change})} \right\} \bullet \frac{P}{Q}.$$

## Elasticity and a Linear Demand Curve

The slope of a straight line demand curve is constant; the inverse of a constant slope is constant. The first term in the elasticity coefficient for a straight line demand curve is, therefore, the same for any price/quantity combination along the curve. The ratio of price to quantity varies along the curve, however. Near the vertical axis, price is high and quantity is low, and $\frac{P}{Q}$ is large. Near the horizontal axis, price is low and quantity is high, resulting in a low ratio. When $\frac{P}{Q}$ is multiplied by the absolute value of the constant inverse of the slope, the resulting elasticity varies along the demand curve. At high prices, elasticity is high, and at low prices, it is low. A linear demand curve has the algebraic form

$$P = a - b \bullet Q,$$

where *a* and *b* are constants.

The price elasticity of demand for a linear demand curve is a function of price and the vertical intercept, shown by

$$\varepsilon_p^d = \frac{1}{b} \bullet \frac{P}{Q} = -\frac{1}{b} \bullet \frac{P}{\frac{(a-P)}{b}} = (-1) \bullet \frac{P}{a-P}.$$

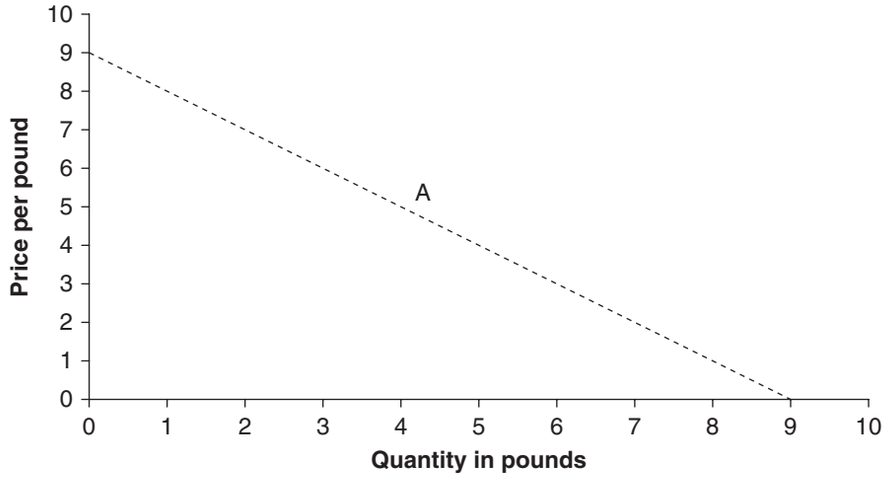


**Figure 9.1** Demand for Beef

The demand curve is unit elastic where

$$\varepsilon_p^d \quad \left\{\frac{1}{(\text{slope of demand curve for given price change})}\right\} \bullet \frac{P}{Q} \qquad 1.$$

$$\varepsilon_p^d = -\left(\frac{1}{b} \bullet \frac{P}{Q}\right) = -1.$$

Because the inverse demand is $Q = \frac{a}{b} - \frac{P}{b}$, at unit elasticity,

$$\frac{1}{b} \bullet \frac{P}{\left(\frac{a}{b}\right) - \left(\frac{P}{b}\right)} = \frac{1}{b} \bullet \frac{P}{\left(\frac{1}{b}\right) \bullet (a - P)} = 1$$

$$\frac{P}{a - P} = 1,$$

$$P = a - P,$$

$$2 \bullet P = a,$$

$$P = \frac{a}{2}.$$

Unit elasticity occurs where price is equal to half the vertical intercept on the demand curve. Because the demand curve is linear, this occurs at the midpoint of the curve (see Figure 9.2).

Flatter straight-line demand curves are described as being more elastic than steeper ones. Straight-line curves, however, have all ranges of elasticity. How can these be reconciled? If a relatively flat straight-line demand is extended until it hits both axes, it will be elastic in the upper half of its price range and inelastic in the lower half of its price range. Similarly, if a relatively steep straight-line demand curve is extended until it hits both axes, it will be elastic in the upper half of the price range and inelastic in the lower half. If a relatively steep demand curve, $D_1$, intersects a relatively flat demand curve, $D_2$, at a given point ($A$ in Figure 9.3), the curve that is steeper will have a lower elasticity at that point than the relatively flatter demand curve. Both curves have the same coordinates ($Q_A$, $P_A$), and the reciprocal of the demand curve is lower for the steeper curve than for the flatter curve. Therefore,

$$\varepsilon = \frac{dQ}{dP} \bullet \frac{P}{Q}$$

will be greater in absolute value for the flatter curve than for the steeper demand curve.

## Elasticity of Demand for Nonlinear Demand Curves

Nothing in the economic theory of demand requires that demand curves be linear; the correct form of the demand function depends on the underlying relationship between quantity and willingness to pay. The definition of elasticity can be applied to determine the elasticity coefficient for alternative specifications of the demand function. One frequently used specification of the demand function takes the following log-linear form:

$$Q = \frac{\alpha}{P^{\beta}} = \alpha \bullet P^{-\beta}.$$

where $Q$ is quantity demanded and $P$ is price; $\alpha$ and $\beta$ are parameters.

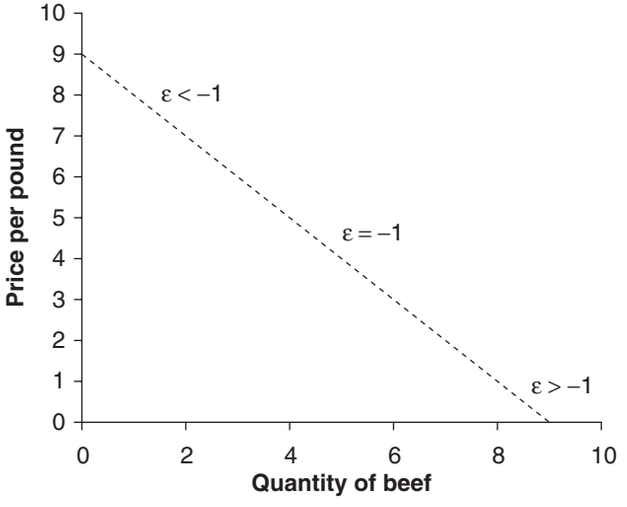


**Figure 9.2** Demand for Beef: Elasticities Along Straight-Line Demand Curve

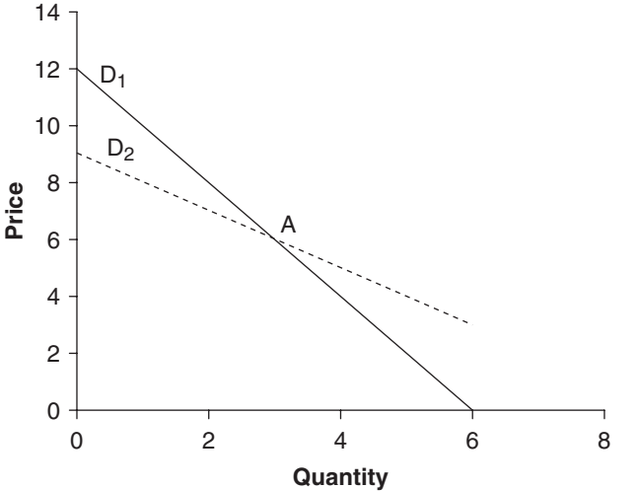


**Figure 9.3** Elasticities of Steep Versus Flat Demand Curves

The log-linear designation derives from the observation that

$$\log Q = \log \alpha + (-\beta) \bullet \log P = \log \alpha - \beta \bullet \log P.$$

Because this function is linear (in the logarithms of the variables), standard linear regression analysis permits the estimation of the parameters, $\alpha$ and $\beta$.

For this log-linear demand function, price elasticity is

$$\begin{aligned} \varepsilon &= \left(-\beta \bullet \alpha \bullet P^{-\beta-1}\right) \bullet \frac{P}{Q} \\ &= \left(-\beta \bullet \alpha \bullet p^{-\beta-1}\right) * \frac{P}{\alpha \bullet P^{-\beta}} \\ &= \frac{-\beta \bullet \alpha \bullet P^{-\beta}}{\alpha \bullet P^{-\beta}} = -\beta. \end{aligned}$$

With a log-linear demand function, the price elasticity of demand is the exponent of price in the demand function. Because the log-linear function specifies the exponent as constant, the price elasticity of demand becomes constant, unvarying as price and quantity change.

Both linear and log-linear demand specifications are relatively tractable; it is relatively easy to derive the price elasticity and relatively easy to use standard regression programs for empirical estimation. One can construct more complex demand functions, with price and quantity inversely related, and derive elasticity using that concept's definition; with greater complexity, however, come elasticities that are messier to calculate and estimate statistically.

## Price Elasticity of Demand and Revenue

Total revenue measures price times quantity for given coordinates ($Q$, $P$) on the demand curve. From the perspective of consumers, total expenditure is the equivalent of total revenue. Along the demand curve, price and quantity move in opposite direction: When $P$ falls, the quantity consumers are willing to buy increases. Change in total revenue (or total expenditures) results from price moving in one direction and quantity moving in the opposite direction and depends on the relative change in the two variables. If the rise in quantity is large relative to the decline in price, the product $P \bullet Q$ rises. Conversely, if the rise in quantity is small relative to the fall in price that induces the quantity change, total revenue (the product $P \bullet Q$) falls. The change in quantity relative to a given change in price is just the price elasticity of demand. If the percentage rise in quantity is large relative to the percentage decline in price, total revenue (expenditure) rises with a fall in price, and the elasticity coefficient is less than −1 (or greater than 1 in absolute value). If a price increase induces a large decrease in quantity demanded, the demand curve is elastic, and total revenue (expenditure on the product) falls. Thus, if demand is price elastic, price and total revenue (expenditure) move in opposite directions: A decrease in price raises total revenue, and an increase in price reduces total revenue.

If a price decrease generates a relatively small increase in quantity demanded, the percentage change in quantity divided by the percentage change in price will be greater than −1 or less than 1 in absolute value. A price decrease leads to a fall in the product $P \bullet Q$ or total revenue. Conversely, if a price increase leads to a relatively small decrease in quantity demanded, total revenue will rise with a rise in price. In this case, the price elasticity is greater than −1 or less than 1 in absolute value. If the demand curve is price inelastic, price and total revenue move in the same direction. When the demand curve has unitary elasticity, the percentage rise in quantity equals the percentage fall in price. The rise in quantity just offsets the fall in price, leaving total revenue unchanged with unitary price elasticity of demand.

Because price elasticity of demand determines the change in total revenue resulting from a price decrease/quantity increase, price elasticity also determines the marginal revenue due to a price/quantity change. Marginal revenue ($MR$) is defined as the change in total revenue resulting from a small change in quantity:

$$MR = \frac{\Delta TR}{\Delta Q}.$$

If the demand curve is elastic for a given price change, total revenue rises with a fall in price and corresponding rise in quantity. Both numerator and denominator of marginal revenue are positive, and therefore marginal revenue is greater than zero. If the demand curve is inelastic for a given price change, total revenue falls with a fall in price and corresponding rise in quantity. The change in total revenue is negative when the change in quantity is positive and the demand is inelastic. Marginal revenue with a price decrease is therefore negative. If the price elasticity of demand is unitary, total revenue is unchanged with a fall in price and corresponding rise in quantity; this results in marginal revenue of zero. Table 9.3 summarizes these relationships.

## Demand Elasticities, Market Power, and Pricing Policies

Figure 9.3 contains two demand curves, one relatively flat; the analysis explained that the flatter curve is more elastic at the intersection of the two than is the steeper curve. A single, small firm in a competitive market faces a horizontal demand curve because the firm has no power over market price; it is a price taker, accepting the market price

**Table 9.3** Price Elasticity of Demand, Total Revenue (Expenditure), and Marginal Revenue

| *Elasticity* | *Absolute Value of Elasticity* | *Condition* | *Total Revenue/ Expenditure* | *Marginal Revenue* |
|---|---|---|---|---|
| $\varepsilon < -1$, demand elastic | $\lvert\varepsilon\rvert > 1$ | Price falls, quantity rises | Rises | $> 0$ |
| $\varepsilon < -1$, demand elastic | $\lvert\varepsilon\rvert > 1$ | Price rises, quantity falls | Falls | $> 0$ (both $\Delta TR$ and $\Delta Q$ are negative, $\frac{\Delta TR}{\Delta Q} > 0$) |
| $\varepsilon = -1$, demand unitary elastic | $\lvert\varepsilon\rvert = 1$ | Price falls, quantity rises | Does not change | $= 0$ |
| $\varepsilon = -1$, demand unitary elastic | $\lvert\varepsilon\rvert = 1$ | Price rises, quantity falls | Does not change | $= 0$ |
| $\varepsilon > -1$, demand inelastic | $\lvert\varepsilon\rvert < 1$ | Price falls, quantity rises | Falls | $< 0$ ($\Delta TR$ is negative and $\Delta Q$ is positive, $\frac{\Delta TR}{\Delta Q} < 0$) |
| $\varepsilon > -1$, demand inelastic | $\lvert\varepsilon\rvert < 1$ | Price rises, quantity falls | Rises | $< 0$ ($\Delta TR$ is positive and $\Delta Q$ is negative, $\frac{\Delta TR}{\Delta Q} < 0$) |

as outside its control. It faces a horizontal demand curve. In this situation, illustrated in Figure 9.4, the price elasticity of demand for the firm is $\frac{\Delta Q}{\Delta P} \bullet \frac{P}{Q} = \frac{\Delta Q}{0} \bullet \frac{P}{Q} = \infty$, or infinitely elastic. At the other extreme is a firm facing a vertical demand curve. The elasticity of demand in this case equals $\frac{0}{\Delta P} \bullet \frac{P}{Q}$, which equals 0 or completely inelastic.

The perfectly competitive firm, which has no market power, faces an infinitely elastic demand curve. More generally, it is possible to link a firm's market power to the elasticity of demand that it faces. Abba Lerner (1934) defined a firm's market power as its ability to raise price above marginal cost, where marginal cost is the additional cost of producing one more unit of output.

$$\text{Lerner Index (of market power)} = \frac{P - MC}{P}.$$

The profit-maximizing firm will produce a quantity such that the addition to revenue equals the addition to cost—that is, marginal revenue equals marginal cost. For the profit-maximizing firm, the index becomes

$$\text{LernerIndex} = \frac{P - MR}{P}.$$

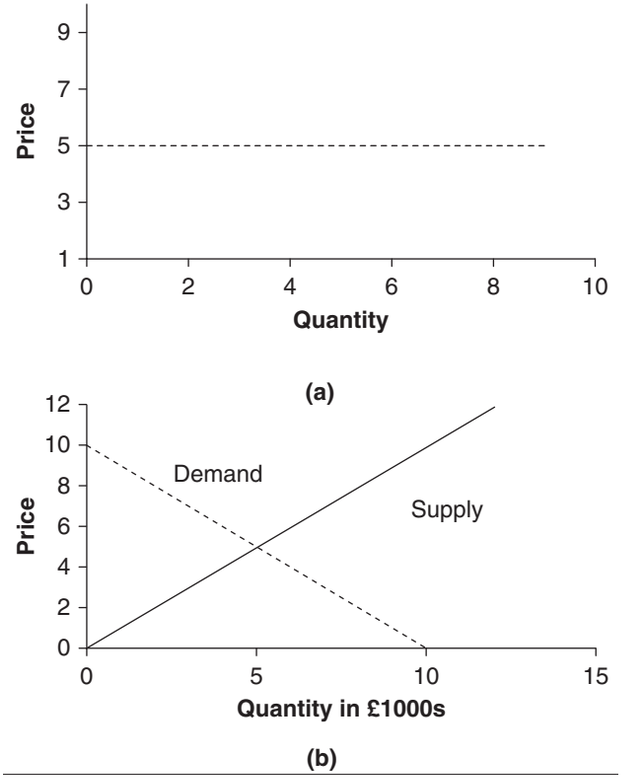


**Figure 9.4** (a) Demand Facing Competitive Firm With Industry Demand; (b) Supply and Equilibrium Price = 5

Total revenue equals price times quantity. Therefore,

$$MR = \frac{dTR}{dQ} = \frac{d(P \bullet Q)}{dQ} = P \bullet \frac{dQ}{dQ} + Q \bullet \frac{dP}{dQ}.$$

Multiplying the right-hand side of this equation by $\frac{P}{P}$ yields

$$MR \quad \frac{P}{P} \bullet \left(P + Q \bullet \frac{dP}{dQ}\right) \quad P \bullet \left(1 + \frac{Q}{P} \bullet \frac{dP}{dQ}\right) \quad P \bullet \left(1 + \frac{1}{\varepsilon}\right).$$

where the elasticity is negative. The Lerner Index of market power becomes

$$\begin{aligned} \text{Lerner Index} &= \frac{P - MR}{P} \\ &= \frac{P - P \bullet \left(1 + \frac{1}{\varepsilon}\right)}{P} \\ &= \frac{P}{P} \bullet \left(1 - 1 - \frac{1}{\varepsilon}\right) \\ &= -\frac{1}{\varepsilon} \end{aligned}$$

If the firm faces a highly elastic demand curve, $\varepsilon$ is large and $-\frac{1}{\varepsilon}$ approaches 0. Therefore, $\left(1 + \frac{1}{\varepsilon}\right)$ approaches 1, *MR* approaches *P*, and again the index approaches 0. If price elasticity of demand is low, $\frac{1}{\varepsilon}$ becomes large, and *MR* diverges from price. The Lerner Index of market power rises.

### Price Elasticity of Demand and Price Discrimination

Price discrimination occurs when different customers for whom the marginal cost of the good or service is the same are nonetheless charged different prices. An example familiar to all tuition-paying undergraduates is the service of college education. Although the list or stated price for a semester at a given school typically is the same for all registrants, financial aid programs effectively create different payments for what is essentially the same service that arguably has the same marginal cost for all consumers. For a seller to price discriminate, that seller has to have market power. Different consumers or groups of consumers must have different demands for the goods, and the seller must also be able to distinguish those willing and able to pay high prices from those willing or able to buy only at low prices. Furthermore, resale of the good or service must be costly or impossible; otherwise, low-price payers would resell to high-price buyers. Under these circumstances, the seller will find it advantageous to separate the markets, selling at a high price in the market with low price elasticity of demand and at low(er) price in the market with a high(er) price elasticity of demand. Airlines attempt to distinguish business from leisure travelers. The former, at least in boom times, are less sensitive to price of the airline ticket either because the firm is paying for it or because the financial benefits resulting from travel exceed the cost of the ticket. No business flyer has reason, however, to disclose a willingness to pay a high price. The airlines therefore set up fare structures that lead buyers to identify themselves as price insensitive or price sensitive. One such structure has lower fares for trips that include a Saturday night stay in the destination city than trips that have returns on weekdays. Presumably business travelers want to be home by the weekend, but leisure travelers are willing to "stay over" for a lower fare. Similarly, retail stores use coupons and rewards cards to distinguish those who are sensitive to price from those who are not. Carrying coupons and cards takes some time and attention, costly to those who value time and convenience more highly than they do the savings from lower price. Price discrimination can increase market efficiency in the sense of leading to higher output than would occur when a single price firm has market power. Also, charging different consumers different prices based on differing price elasticities of demand generates greater efficiency when fixed costs are high relative to variable ones—or when some costs are not easily adjusted for small changes in output. Economists interested in antitrust and regulation have become very interested in studying pricing schemes based on differences in price elasticity.

Because lower price elasticity of demand allows a firm to charge higher prices, ceteris paribus, a firm has an incentive to reduce that elasticity. One way of doing this is to generate consumer loyalty. Developing a brand name presumably does this, as does some degree of product differentiation. In these cases, if price rises, not all customers flee to alternative products. Advertising potentially increases brand loyalty and reduces price elasticity of demand.

## Other Demand Elasticities

### Income Elasticity of Demand

This chapter has focused thus far on the relationship between the price of a good or service and the quantity demanded at that price. Clearly, factors other than price of a good affect the willingness of consumers to purchase. More generally, the demand for good *x* is a function of the price of *x,* income (*M*), and price of a related good, *y.* Of

course, there may be more than one related good whose price affects the demand for *x*.

$$Q_x^d = f(P_x, M, P_y).$$

Generally, an elasticity is the percentage change in the dependent variable divided by the percentage change in the independent or causal variable, holding all other relevant variables constant. When the price of good *x* is held constant and income is allowed to vary, the resulting relationship between income and the quantity of good *x* that consumers are willing to buy is graphed as an Engel curve, as shown in Figure 9.5.

The arc income elasticity of demand along the Engel curve is

$$\varepsilon_M^d = \frac{\Delta Q}{\Delta M} \bullet \frac{\sum M}{\sum Q}.$$

The point income elasticity of demand is

$$\varepsilon_M^d = \frac{dQ}{dM} \bullet \frac{M}{Q}.$$

The first term of the elasticity is the marginal propensity to consume the good, that is, the change in the consumption of the good divided by the change in income. The second term in the point elasticity is the reciprocal of the average propensity to consume the good. Thus, the income elasticity of demand equals *MPC* ÷ *APC* for the good. At a given point, the *MPC* is the slope of the Engel curve. The *APC* is the slope of a straight line, a ray, from the origin to the given point on the Engel curve. The income elasticity of demand equals the ratio of the slope of the Engel curve at a given point to the slope of a ray from the origin to that same point. If the Engel curve is steeper than the ray from the origin, the income elasticity at that point exceeds 1. If the Engel curve is less steep than the ray from the origin at the point of interest, the income elasticity of demand at that point is less than 1.

The sign of the income elasticity provides important information. When its income elasticity of demand is positive, a good is called normal; a rise in income leads to a rise in the demand for the good, a rightward shift of the demand curve. The Engel curve is positively sloped. A good is called inferior when its income elasticity of demand is negative. When income rises, the consumption of the good falls; the demand curve for the good shifts to the left with a rise in income, and the Engel curve is negatively sloped. A good is not inherently inferior. At low levels of income, a rise in income may lead consumers to buy more low-grade hamburger. If income rises further, consumers may choose to buy less low-grade hamburger and instead buy ground sirloin.

If a good's income elasticity of demand is positive and high, greater than 1, the good is superior. In the 1950s in the United States, new housing was arguably a superior good. As incomes rose, people spent disproportionately on new housing; new housing was hard to find in city centers, and therefore the demand for suburban housing rose. Government transportation policy favored expansion of highways that connected suburban housing areas to urban and, later, suburban work centers. Urban public policy also favored separation of housing from nonhousing land use. With rising incomes, the high-income elasticity of demand for new housing combined with public policy arguably led to suburban sprawl and accelerated the decline of central cities.

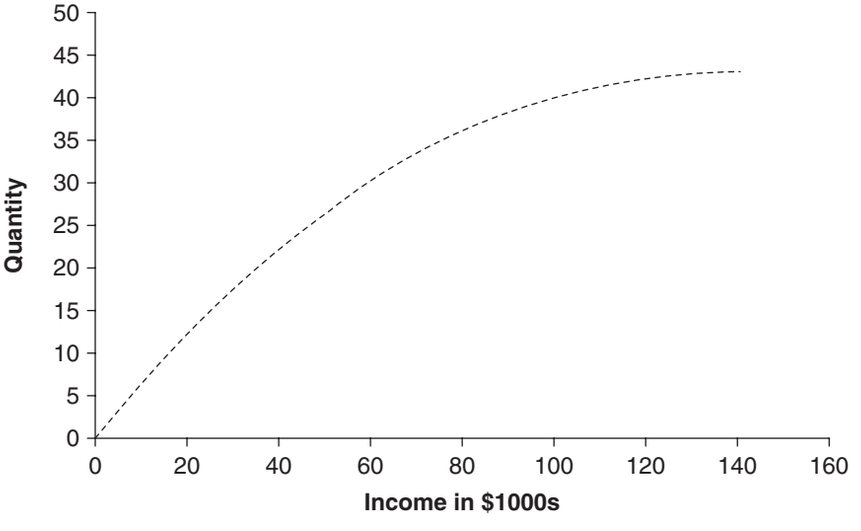


**Figure 9.5** Engel Curve

## Cross-Elasticity of Demand

The cross-elasticity of demand measures the responsiveness of demand for good *x* to the change in the price of good *y:*

$$\varepsilon_{Py}^d = \frac{\Delta Q_x}{\Delta P_y} \bullet \frac{\sum P_y}{\sum Q_x}$$

or for a point cross-elasticity of demand

$$\varepsilon_{Py}^d = \frac{dQ_x}{dP_y} \bullet \frac{P_y}{Q_x}.$$

The sign of the cross-elasticity provides useful information. If the cross-elasticity is negative, a fall in the price of good *y* causes a rise in the quantity of good *x* via a rightward shift of the demand curve for *x.* Consumers are purchasing more *x* when the price of *y* is falling. They are buying more *x* when they buy more *y*; *x* and *y* are complements. If the cross-elasticity is positive, a fall in the price of good *y* leads to a fall in the demand for *x.* The

**Table 9.4** Other Demand Elasticities

| *Type of Elasticity* | *Value of the Elasticity* | *Implications for Good x* |
|---|---|---|
| Income | $\varepsilon > 1$ | Superior |
| Income | $0 < \varepsilon \leq 1$ | Normal |
| Income | $\varepsilon < 0$ | Inferior |
| Cross | $\varepsilon > 0$ | Substitutes |
| Cross | $\varepsilon = 0$ | Unrelated |
| Cross | $\varepsilon < 0$ | Complements |

demand curve for $x$ shifts leftward with a fall in the price of $y$. Consumers substitute $y$ for $x$ when the price of $y$ falls; $x$ and $y$ are substitutes. Table 9.4 summarizes these demand elasticities.

If the demand for good $x$ is log-linear, that is,

$$Q = p_x^{\alpha} \bullet M^{\beta} \bullet P_y^{\delta},$$

the demand elasticities are the exponents of their respective variables. The "own" price elasticity of demand is $\alpha$, the income elasticity of demand is $\beta$, and the cross-elasticity of demand is $\delta$.

## Factors Influencing Price Elasticity of Demand

The modern demand curve for good $x$ is drawn assuming that income and prices of all other goods are constant; it illustrates how a change in the price of $x$ affects the quantity demanded, ceteris paribus, or "all other things constant." The calculation of price elasticity of demand for a price change is based on the same assumption. A change in the price of good $x$, with nominal income held constant, changes the real income or purchasing power of the consumer(s). Hence, along the typical demand curve, real income varies, although one can create an income-compensated demand curve, along which real income is held constant. Along an uncompensated (the more typical) demand curve, real income varies with changes in the price of $x$. If purchases of $x$ represent only a small portion of the consumer's budget, changes in the price of $x$ do not affect real income significantly. If instead purchases of $x$ represent a large portion of the consumer's income, real income effects are substantial and affect the value of the calculated price elasticity of demand, holding nominal income constant. One significant factor in determining the price elasticity of demand is the portion of the budget that is devoted to the good. The smaller the percentage of income spent on good $x$, the more inelastic the demand for the good, ceteris paribus. A doubling of the price of salt in affluent communities is unlikely to reduce consumption noticeably.

Related to this budget share effect is whether the good is normal or inferior. If the good is inferior, the real income effect of a price fall leads to a reduction in the amount of the good purchased. For most goods, this income effect is more than offset by a substitution effect, whereby the quantity of $x$ increases as it is substituted for related goods. The quantity response to a price change is lessened by the income effect, and hence price elasticity of demand is reduced as compared with normal goods.

The number of substitutes for good $x$ strongly affects the price elasticity of demand for good $x$. If $x$ has many close substitutes, consumers will shift to these substitutes when $x$'s price rises. The more narrowly a product group is defined, the higher the price elasticity. For example, the demand for fresh green peas is more elastic than the demand for fresh vegetables, which is more elastic than the demand for all vegetables (fresh, frozen, canned), which is more elastic than the demand for food. Similarly, the demand for a Toyota Prius is more elastic than demand for all Toyota automobiles; the demand for Toyotas is more sensitive to price than the demand for all motor vehicles. Even with goods for which there is no substitute (e.g., insulin for a diabetic), the real income effect of a price increase can lead to decreased consumption if prices become high enough. In lower income countries, it is not unusual for consumers to buy antibiotics one or two pills at a time because prices are high relative to money income.

A fourth factor affecting price elasticity of demand is the length of time under consideration. The longer the time period of adjustment, the more elasticity of demand is expected to be. When the price of gasoline rose in mid-2008, car owners did not immediately sell their low gasoline mileage cars for gasoline-electric hybrids or immediately sell their suburban homes to move close to public transportation. They did reduce the number of miles driven, however, but not in as great a proportion as the rise in the price of gasoline. Had prices remained high (and macroeconomic economic conditions not worsened

significantly), consumers would have adjusted over time perhaps by buying more fuel-efficient cars, considering telecommuting to work periodically, moving closer to work, or commuting with others. For these more substantial adjustments to be made, prices do have to remain high and consumers must anticipate that they will remain high. In the long run, demand is more price elastic than in the short run.

## Elasticity of Demand for an Input

Building on the work of both his classical and marginalist predecessors, Alfred Marshall (1920) considered the responsiveness of input demand to the change in the price of an input to the production process. In this he was considering joint production, where the input is one of several in the production of the output. The demand for an input is a derived demand, derived from the demand for the good that the input produces. Marshall concluded four conditions affected the elasticity of demand for the input: how essential the input was to the production of the final product, the elasticity of demand for the final product, the fraction of total cost accounted for by the input, and the elasticity of supply of cooperating inputs. The more essential the input, the lower the elasticity of demand for it. The lower the elasticity of demand for the product, the lower the price elasticity of demand for the input. The lower the fraction of total costs that the input represents, the lower the price elasticity of demand for the input. The more inelastic the supply of cooperating inputs, the lower will be the price elasticity of demand for the given input.

When contemplating an increase in the minimum wage, policy makers are concerned about the elasticity of demand for low-skill labor. If that elasticity is low, a rise in the wage will not decrease the quantity of labor demanded by much. The rising legal floor under the wage rate will not lead to a large increase in unemployment, given a relatively low elasticity of labor supply. If the demand for low-skill labor is very elastic, policy makers should be concerned about the potential impact of the rising legal wage on employment.

Labor economists are concerned about the impact on labor markets of the retirement of large numbers of baby boomers. If the demand for labor is elastic, the falling supply of labor will raise wages, but not dramatically. If the demand is wage inelastic, wages are likely to rise much more, ceteris paribus. The ease of substituting capital for labor partly determines the elasticity of demand for labor. If the supply of capital is elastic, the demand for labor will be more elastic as firms find that they can hire capital at relatively stable prices to replace the retiring boomers.

## Policy Implications

Elasticities of demand influence the quantitative impact of public policies, and hence applied and public policy microeconomists use them frequently. One important application involves the incidence of a tax. The incidence of a tax is the answer to the question, "Who pays the tax?" If the tax is levied on the seller, the equilibrium after-tax price of the product will rise and equilibrium quantity will fall. The net price to the producer must fall because the tax must be paid out of the new equilibrium price, unless price rises by the full amount of the tax. If the inverse demand for the good is $P = a - b \bullet Q$ and the inverse supply is $P = c + d \bullet Q$, the proportion of the tax paid by the consumer will equal to $b/(b + d)$. The coefficients $b$ and $d$ are the slopes of the demand and supply curves, respectively, and hence enter into the elasticities of demand and supply. Applying the definition of elasticity to these linear curves, we get the result that the proportion of the tax paid by the consumer equals $(-1) \bullet \left( \frac{1}{\varepsilon_d - \varepsilon_s} \right)$. If the demand elasticity is large, the proportion paid by the consumer is small. If the elasticity of supply is large, relative to the elasticity of demand, the consumer pays a larger proportion of the tax.

If the demand for the good is price inelastic, the effect of the tax will be higher price but not much reduction in the quantity consumed. If the purpose of the tax is to raise revenue, the tax on a good whose demand is price inelastic will be successful. Total tax collected is a function of the tax rate (either specific amount per unit of the good or ad valorem, a percentage of the good's price) times the quantity sold. If quantity does not decline much as a result of the tax, total tax collections are large. If the intent of the tax is to raise revenues, it should be levied on a good with price-inelastic demand. If demand is price elastic, equilibrium quantity will fall significantly as a result of the tax and higher price; tax revenues will be low. Policy makers seeking to discourage consumption of a particular good (e.g., cigarettes) can levy a heavy tax on the good. If demand is price inelastic, however, consumption will not fall dramatically. Some who propose hefty cigarette taxes to deter consumption argue that the real income effect of the tax increase will fall heavily upon young smokers, discouraging potential smokers before they become older income-earning addicted smokers. Older smokers, for whom the real income effect is less constraining because they have higher incomes, will continue smoking due to addiction, addiction that suggests that there are no close substitutes for cigarettes. Tax supporters argue that while the short-term effect of high cigarette taxes will be high tax revenues, in the longer run, such taxes will deter smoking.

Economists have used the elasticity concept to analyze perceived problems in many sectors of the economy,

including agriculture. They note that the demand for agricultural goods tends to be price inelastic. Large increases in agricultural productivity that increase supply lead to lower prices, and lower prices lead to lower farm revenues. The demand for agricultural goods is income inelastic in high-income countries. The demand for basic foodstuff therefore rises less than proportionately to income. As a result of both price and income inelasticity, prices of farm goods are likely to fall over time relative to the prices of nonfarm goods, ceteris paribus.

## Empirical Estimates

R. A. Lehfeldt (1914), in an early attempt to measure price elasticity of demand, examined the wheat market and found that defining the product was problematic due to the wide variation in quality. Lehfeldt impressively overcame problems of market definition and estimated the elasticity to be .6 (absolute value), but with many a caveat.

To estimate demand and demand elasticities, the empirical economist has to find relevant data. Many demand elasticity estimates focus on agricultural goods. The U.S. Department of Agriculture was established in 1862, with one of its functions being the collection and dissemination of data. There exists, therefore, a substantial data set and a policy-oriented market for studies on agricultural goods. Regulation of an industry also tends to generate publicly available data. Telecommunications and electricity were heavily regulated at national, state, and local levels from the 1930s through the 1980s and still face regulatory constraints. Both the available data and the debates over pricing schemes for regulated industries generated numerous demand and elasticity studies. For both industries, the impact of time-of-day pricing on peak load use involved elasticity considerations: If usage were highly sensitive to price, lower prices in off-peak hours could reduce peak-load capacity needs. Another public utility, water provision, is often studied for demand sensitivity to price. Government regulation generates data on liquor prices and sales, reflected in the numerous studies of elasticities of liquor demand. The same is true for cigarettes and for legal gambling. Studies of the income, price, and interest rate elasticities of demand for housing have appeared, related to the impact of this sector on the well-being of the poor, the economic condition of cities, and macroeconomic stability concerns. Mid-twentieth-century studies of transportation tended to focus on the income elasticity of demand for private versus public transportation. More recently, transportation studies have related to the need for energy conservation. The periodic "energy crises" from the 1970s to the present have stimulated price–quantity–elasticity studies of energy use, oil demand, gasoline use, and use of alternative fuels. Economists also tried to estimate price and income elasticities of demand for health insurance and health care. Many studies have sought to measure the elasticity of demand for labor, often in conjunction with predictions about the effect of a change in labor legislation. Of course, researchers have undertaken elasticity studies on many other goods and services, but availability of data and pertinence to public policy issues seem important in choice of study area.

When Marshall (1920) refined the definition of elasticity, access to computational tools limited its actual measurement. The advent and spread of high-speed computers, development of statistics software packages that eliminate tedious calculation, and the availability of electronic data sets facilitate greatly data collection and calculations. In recent years, the use of scanners by retail outlets to record price and sales has generated a potentially large data set. Although usually proprietary, such data are selectively available (e.g., in antitrust cases concerning mergers) and have been used, for example, to consider cross-elasticities of demand between alternative sources of office supplies.

As economists have noted, the concepts of price, income, and cross-elasticity are relatively simple and easy to understand. Measurement of actual elasticities quickly seems less simple. Problems occur in defining the product, in determining the appropriate time period, in determining the best functional form for the demand function, and in estimation procedures. Most daunting is that demand curves are constantly shifting, making it difficult to sort out movements along a demand curve from a shift in the demand. Nonetheless, elasticity estimates are essential to understanding economic change and to good policy making.

## Note

1. Keynes (1963) reports in a footnote that Mrs. Marshall told him that her husband had hit upon the notion of elasticity while on the roof in Palermo. His excited reaction may be pure legend.

## References and Further Readings

Adams, W., & Brock, J. (2005). *The structure of American industry* (11th ed.). Upper Saddle River, NJ: Pearson/Prentice Hall.

Besanko, D., & Braeutigam, R. (2008). *Microeconomics* (3rd ed.). Hoboken, NJ: John Wiley.

Ekelund, R. B., Jr., & Hebert, R. (1990). *A history of economic theory and method* (3rd ed.). New York: McGraw-Hill.

Frank, R. (2004). *Microeconomics and behavior* (6th ed.). Boston: McGraw-Hill Irwin.

Friedman, M. (1962). *Price theory: A provisional text.* Chicago: Aldine.

Henderson, J., & Quandt, R. (1980). *Microeconomic theory: A mathematical approach* (3rd ed.). New York: McGraw-Hill.

Keynes, J. M. (1963). *Essays in biography* (G. Keynes, Ed.). New York: W. W. Norton.

Kreps, D. M. (1990). *A course in microeconomic theory.* Princeton, NJ: Princeton University Press.

Lehfeldt, R. A. (1914). The elasticity of demand for wheat. *The Economic Journal, 24,* 212–217.

Lerner, A. (1934). The concept of monopoly and the measurement of monopoly power. *Review of Economic Studies, 1*(3), 157–175.

Marshall, A. (1920). *Principles of economics* (8th ed.). London: Macmillan.

McCloskey, D. N. (1985). *The applied theory of price* (2nd ed.). New York: Macmillan.

Raffaelli, T., Becattini, G., & Dardi, M. (Eds.). (2006). *The Elgar companion to Alfred Marshall.* Cheltenham, UK: Edward Elgar.

Schumpeter, J. (1965). *Ten great economists from Marx to Keynes.* New York: Oxford University Press.

Varian, H. (2006). *Intermediate microeconomics: A modern approach* (7th ed.). New York: W. W. Norton.

# 10

# Costs of Production

## *Short Run and Long Run*

Laurence Miners
*Fairfield University*

When most people other than economists think of costs, they may logically think about their household budget and the cost of heating their home or the cost of sending a daughter or son to college. Economists, however, are more apt to talk about the *prices* of these items and consider how households allocate a finite income to meet family needs. When economists consider *costs,* they refer most often to the production decisions of firms—what and how much they decide to produce, how they produce it, and how much it costs. The usual free-market assumption of profit-maximizing behavior by firms is not necessary for this discussion. All firms, from small local nonprofits, such as community libraries, to large international corporations attempt to operate efficiently. That is, they strive to produce the most output at the lowest possible cost.

The purpose of this chapter is to consider the production decisions and associated costs that firms face. It should be clear at the outset that this discussion will be incomplete in that it will not consider the profitability of a firm or the particular market in which it operates. A firm may produce a safe, reliable product at minimum cost, using the best technology, but fail if there is insufficient demand for its product. Similarly, an inefficient, lumbering, pollution-generating company may make significant profits if it dominates its industry and has a loyal following of customers. This is not meant to be seen as an endorsement of any particular industry structure. Rather, it is to point out that the costs and production decisions that firms make address only part of the economic survival equation. One must also consider the demand for the firm's product and the market in which it operates. (The remaining chapters in this section, as well as many of those in Parts III and VI, address these critical issues.)

The discussion of costs in most introductory and intermediate microeconomic textbooks is theoretical in nature, and the presentation here will likewise be primarily theoretical. The first part of the chapter investigates the relationships among inputs, production, and costs of the firm. The second part of the chapter discusses the relationship between short-run and long-run costs. Following that, we consider empirical estimates of firm output and costs.

## Theory

### The Production Function

As Calvin Coolidge famously remarked, "The business of America is business." Firms come in all shapes and sizes—from small individually owned entrepreneurships to large multinational corporations. However, they all combine inputs to produce products, outputs, or goods and services. When speaking about the production process of a firm, these last expressions—products, outputs, and goods and services—can all be used interchangeably. Firms hire, rent, or purchase inputs to produce their products. Economists use production functions to investigate the way in which firms transform inputs into outputs. This relationship can be stated in general form, indicating that output is some function of inputs, or can be set up as a specific mathematical equation that can be empirically tested by collecting and analyzing data.

*Inputs*

Inputs, which are also referred to as factors of production, are broadly classified into three categories: land, labor, and capital. Land includes the property on which businesses are built, farm land, and usually the minerals and natural resources removed from the land itself. The workers employed by a firm are referred to as labor, and at least at the introductory level, these workers are assumed to be homogeneous. That is, Worker A is no different from Worker B, and workers are just as productive in their eighth hour of work as they were in their first. These simplifying assumptions allow the discussion to focus more clearly on the important relationships among inputs, outputs, and costs and do not alter the qualitative results of the model. Capital has special meaning for economists. First, capital is not money. Capital is the result of some previous production process and is typically a piece of durable equipment or machinery that a firm uses in its own production of goods and services. As examples, a shipping company may purchase trucks from a truck manufacturer; a university may purchase computers for its faculty and staff to use. It should be pointed out that the specific use of the equipment has an important bearing on how commodities are classified. If a homeowner purchases a truck to take her household refuse to the town dump, that truck is not considered capital; it is a consumer durable. Similarly, if you buy a computer so you can e-mail and send photos to your family and friends, it would not be considered a capital good. The piece of equipment must be used in the production of other goods and services, which are then sold, for it to be considered capital. The amount of money firms spend in producing or procuring capital is called investment spending or, simply, investment. People who work in financial markets often talk about their investments or the amount of capital they have invested in the (financial) market. Economists prefer to refer to these activities as *financial* investments and are consistent in their use of capital as a physical factor of production.

As another example of capital, human capital refers to the investments individuals make in themselves via education and training to become more productive workers. There is a strong analogy between physical and human capital. Both individuals and firms engage in investment spending to acquire and expand their stock of capital. Both types of capital deteriorate over time but can be renewed with additional investment expenditures. Entrepreneurial ability, a special type of human capital, is often considered a fourth factor of production. The entrepreneur assumes the risk in running a business and provides the innovation needed to effectively use the other factors of production.

Finally, firms also use a great deal of intermediate products when they produce goods and services. Intermediate products are themselves outputs of some previous production process and are then put to use in the production of some final good or service. A product is considered final when it is purchased and used by the end user or consumer. Intermediate goods get used up during the production process. For example, a construction company uses lumber when it builds a new home. The primary factors of production—land, labor, and capital—are typically not used up during the production process. To be sure, their economic lives are finite, and they deteriorate, get worn out, and depreciate, but they can be used repeatedly to produce output. Workers, machines, and the land can be used season after season, year after year, to produce products. Lumber and other intermediate products get used up during the production process. To produce additional homes, the construction company must get more lumber; it may not need to hire additional workers or purchase new capital equipment. One final word of caution is in order. Like the trucks and computers discussed previously, how the lumber is used ultimately determines how it is classified. If a homeowner purchases lumber to repair his garage, the lumber is considered a final good, not an intermediate product.

*Short- and Long-Run Output*

The production decisions that firms make can be divided into two distinct time periods: the short run and the long run. In the short run, some of the firm's inputs are fixed in quantity and cannot be changed. For most firms, labor is the factor of production that is most easily changed, and land and capital are less variable. The long run is often viewed as a planning horizon, and it is during this time that all production decisions can be modified. Firms get created and enter industries in the long run. Similarly, companies go out of business, liquidate all their assets, and exit from industries in the long run. There is no set length of time that separates the short from the long run; it depends on the nature of the firm and its production process. For a small corner convenience store, the long run may be months. For a large corporation, with many factories, it may be years. The discussion immediately below focuses on the short run; the long run will be considered subsequently.

*Marginal Analysis*

To explain the relationship between inputs and output, assume that the firm varies one input at a time and holds all its other inputs constant. All firms need to know how productive their inputs are. That is, they need to know how much an additional input will add to output (and revenue) and how much it will cost to hire (or purchase) that input. This so-called marginal analysis, or decision making at the margin, lies at the heart of all economic decisions. Firms, as well as consumers, are always dealing with finite budgets and asking themselves, "Is it worth it?" Economists use the phrase, *ceteris paribus,* or all other things unchanged, when discussing decisions made at the margin. It is convenient, therefore, to consider a short run with only one variable input.

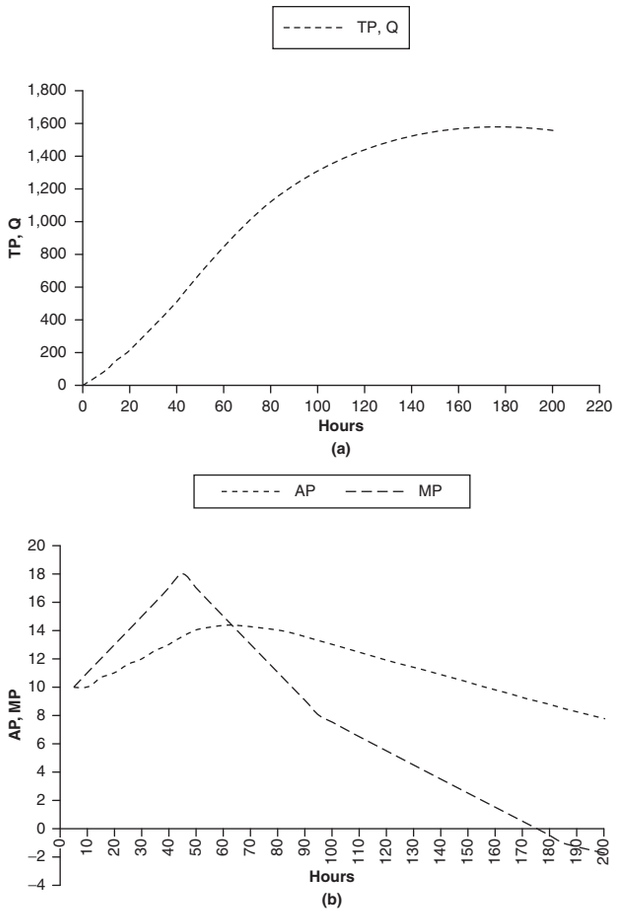


**Figure 10.1** Short-Run Output–One Variable Input: (a) Total Product; (b) Marginal and Average Product

Consider panel (a) in Figure 10.1. This diagram illustrates the firm's quantity of output ($Q$) or total product ($TP$) in the short run, with one variable input, labor. Labor is assumed to be homogeneous and is measured as the hours of labor that the firm uses with its other fixed factors of production. Note that for the most part, as the firm uses more hours of labor, output increases. At some point, however, remembering that this is the short run and other factors of production are held constant, output levels off and even decreases as the firm keeps adding more and more hours of labor. The workers simply run out of capital equipment and other factors that they can use effectively. In this diagram, this appears to happen after about 175 hours of labor.

More interesting is what happens as the firm first begins to use additional workers. Total product increases at an increasing rate. That is, the slope of the $TP$, $Q$ line gets steeper—up to a certain point. Consider a firm that removes baggage from an airplane. The firm uses some capital equipment such as the conveyor belt that is driven up to the side of the plane and the cart used to transport the luggage into the terminal. In the short run, if there is only one worker, the worker needs to get into the plane, place the luggage on the conveyor belt, and then meet it on the ground and place it on the cart. With two workers, one can stay in the plane and the other can stay on the ground. Two workers, working for 1 hour, do more than twice the work of one worker working for 2 hours alone. The marginal product, or contribution to output of the second worker, is greater than that of the first. Marginal product is measured as the change in total product divided by the change in variable input or, in this case, the change in bags removed over the change in hours of labor. Marginal product increases and the $TP$ line gets steeper, at least initially. Eventually, the marginal product has to get smaller. Remember, this is the short run, and the other factors of production are being held constant. Additions to the variable input, in this case labor, cannot keep getting more and more productive. Each additional worker has less of the fixed inputs to work with. This concept is so fundamental to the production process that economists refer to it as the law of diminishing returns or the law of diminishing marginal product. This is not a bad thing; it is both normal and natural. In panel (a) of Figure 10.1, it appears that diminishing returns set in after about 45 hours of labor. Additional hours of work still contribute positively to output. It is just that now, those marginal contributions are getting smaller, and the slope of the $TP$, $Q$ line gets flatter.

### *Marginal and Average Product*

Panel (b) in Figure 10.1 illustrates the marginal and average product curves that are associated with the total product curve in panel (a). Here, it is much easier to see where the law of diminishing returns becomes effective. Marginal product peaks at 45 hours of labor and then starts decreasing. Total product is steepest at this point and this coincides with the peak of marginal product. It is here that diminishing returns set in. As labor hours continue to increase, the marginal product of labor remains positive until about 175 hours of labor. Throughout, marginal product measures *changes* in total product as labor is increased (or decreased) by small amounts. This is precisely what the slope of the total product curve measures as well. The marginal product curve is a plot of the slopes of total product. After 175 hours of labor, total product begins to decline; the slope of the total product curve and marginal product itself become negative at that point.

Measuring average product is fairly straightforward. Like any average, it is calculated by adding up the total number of units, in this case the number of bags removed from the plane, and dividing by the number of hours (of labor). Of special interest is the relationship between average and marginal product. Note that when average product is increasing, the marginal product lies above it (although the margin itself may be decreasing). When the average is

falling, the margin lies beneath it. And when the average is flat (in the vicinity of its peak), the margin crosses the average and is equal to it.

This relationship between marginal and average is not unique to the product curves. As we will see, this relationship carries over to the cost curves and, for that matter, any other data set for which averages and marginals can be calculated. Think of the class average on an exam. If a new student takes the exam (this would be the change or marginal) and scores higher than the average, the average will be pulled up, though not all the way up to the marginal grade. Similarly, if a new student scores lower than the class average, this will bring the average down. Finally, if a new student scores exactly at the class average, the average will remain unchanged.

*Short-Run Costs*

A firm's costs are directly related to its ability to produce goods and services. Consider Figure 10.2, panel (a). In the short run, a firm's total costs are separated into fixed and variable costs. Fixed costs are associated with inputs and other items that are held constant and do not vary with output. Examples of fixed costs include the lease or rental fee a firm may pay for the use of capital equipment and the mortgage it owes on property that it owns. The key point is that these costs do not vary with the firm's level of output and must be paid even if the firm produces nothing. Hence, the total fixed cost curve is drawn as a horizontal line. Whether the firm removes 1,500 pieces of luggage from the plane or shuts down and produces nothing, these costs remain fixed—in this case, at about $1,200.

Fixed costs are an important part of determining the profitability of a firm and cannot be ignored, but they are less helpful in determining exactly how much output the firm should produce. For this, marginal cost is much more useful. Just as marginal product measures the incremental changes in output when, ceteris paribus, the utilization of a single input is changed, marginal cost calculates the additional cost associated with producing a little more (or less) output. Like marginal product, marginal cost is the same thing as the slope of its associated total and is based on the following equation: $MC = \Delta TC/\Delta Q$. Because the firm is operating in the short run, we can assume that labor is the firm's only variable input. If the wage rate ($w$) that the firm pays is fixed, or if the firm's effect on the wage rate is negligible, as would be the case if the firm were one of many firms hiring workers, the numerator of the marginal cost equation can be written as $\Delta TC = w(\Delta L)$. That is, in the short run, the firm's costs change as it uses more, or less, of the variable input, labor. Hence, the marginal cost equation can be rewritten as $MC = w\ (\Delta L)/\Delta Q$. The latter part of this equation, $\Delta L/\Delta Q$, is the inverse, or reciprocal, of marginal product. Therefore, $MC = w/MP$. Now the relationship between marginal cost and marginal product is very clear. When marginal product increases, marginal cost decreases and vice versa. It is logical that if the marginal product of an input is increasing—each additional hour of labor is adding more and more to output—the marginal cost of producing additional units of output must be falling.

Look at panel (b) in Figure 10.2. Recall that the marginal product curve looks like a hill; marginal cost looks like a "U." The two curves are mirror images of each other. Note, too, the relationship between marginal cost and both average variable and average total costs. When the margin is less than or below the average, it pulls the average down. When the averages are increasing, the margin must lie above them and pulls them up. Because fixed costs do not vary with output, the concept of a marginal or change does not apply to its shape. Like any average, average fixed cost (*AFC*) equals total fixed cost (*TFC*) divided by a number—in this case, the number of units of output. Or, $AFC = TFC/Q$. Because the numerator is a constant, as *TFC* gets divided by larger and larger units of output, average fixed costs becomes smaller and smaller.

Just as the sum of variable costs and fixed costs equals total cost, average fixed costs plus average variable costs equal average total costs. These equations are written as $TC = TVC + TFC$ and $ATC = AVC + AFC$. This last equation can be rewritten as $TC/Q = TVC/Q + TFC/Q$.

## The Shut-Down Decision

All firms, regardless of their size or the market in which they operate, need to decide how much product to produce. If the demand for a firm's output is weak, and costs exceed revenues, it may eventually decide to liquidate its assets and go out of business. This process takes time, and *exiting* from the industry occurs in the long run. Indeed, the length of time it takes a firm to act on this decision helps define just how long the long run is for firms in a particular industry. In the short run, exiting, by definition, is not an option. The firm can, however, decide to *shut down.* The firm is still very much a part of the industry. It has simply decided, for a short period of time, to produce nothing. During this period of time, the firm can send the workers home and reduce the utilization of any other variable inputs to zero. However, this firm must still pay its fixed costs—the mortgage, interest on loans, and so on. In this unpleasant situation, the firm is no longer deciding how best to maximize revenues or profits but how to minimize losses. The question the firm faces is, "Should it produce and lose money, or shut down and have to cover fixed costs completely out of pocket?" It should be clear that regardless of its decision, the firm has to pay its fixed costs. What the firm needs to determine is the following: If it produces output, will it bring in enough revenue to cover its variable costs? If it does, it can pay the workers (and any other variable inputs) and use any remaining

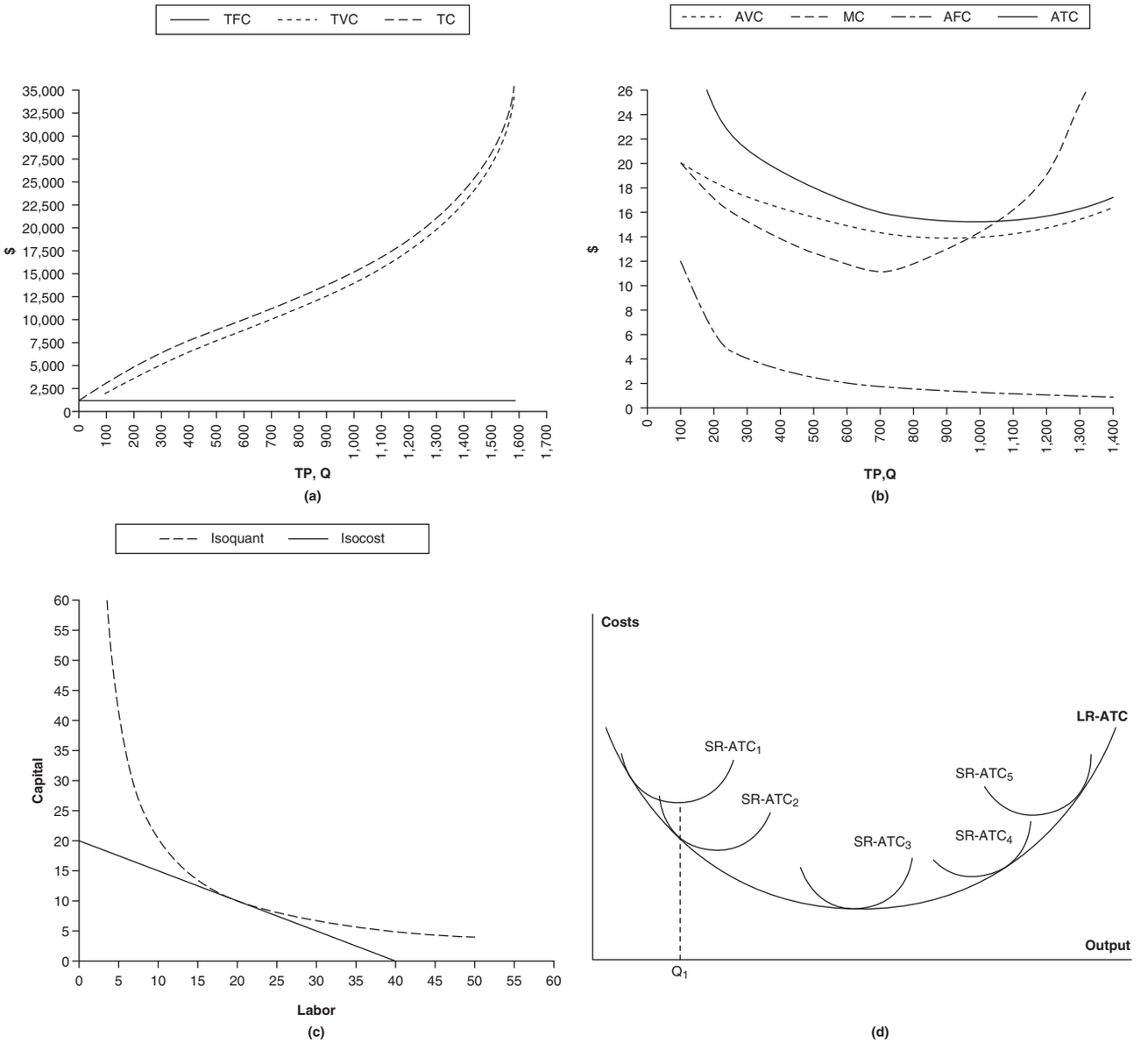


**Figure 10.2** Cost Curves: (a) Total Cost Curves; (b) Marginal and Average Costs; (c) Costs and Output With Two Variable Inputs; (d) Long-Run Average Costs

revenue to defray at least part of fixed costs. Even though the firm will be losing money, this is seen as preferable to shutting down and having to cover all of fixed costs. On the other hand, if the revenues received are not even sufficient to cover the variable costs, why incur them? Send the workers home and just worry about the fixed costs. The short-run shut down point can be seen in panel (b) of Figure 10.2. Once market price falls beneath the minimum of *ATC,* about $15.00, the firm will not bring in enough revenue to cover all of its costs. However, as long as the price remains above the minimum of average variable cost (*AVC*), about $14.00, the firm will be bringing more than enough revenue to pay all of the variable costs, so it will stay open and produce, even though it is losing money. Once the market price falls beneath the minimum of *AVC*, the firm can no longer cover its variable costs and shuts down. Based on this analysis, the short-run supply curve for firms in competitive markets is given as that portion of the marginal cost (*MC*) curve that lies above *AVC* (for a more detailed analysis of this point, see Chapter 11 [this volume]).

While the discussion up to this point has been theoretical in nature, one real-world consideration needs to be

mentioned. Firms generally are reticent to let workers go. Workers who are laid off may find work elsewhere, and when the market improves, the firm will incur the costs of searching for and training new employees. To be sure, firms do lay workers off when their market deteriorates, but the decision is never an easy one.

*Implicit and Explicit Costs*

Until this point, the discussion has focused on actual or explicit expenses that firms incur. Economists, however, are also very concerned about the less indirect or implicit costs that firms face. Implicit costs are associated with any resources that the firm owns and measure the cost of forgone opportunities when resources are used one way instead of another. Implicit costs are associated with capital equipment, retained earnings that the firm has, and the entrepreneurs' allocation of their own time. For example, if a clothing manufacturer owns sewing machines, it may need to decide if this capital equipment should be used to sew shirts or pants. If the firm chooses shirts and makes a profit doing so, it may think it is doing very well. However, the firm should also consider if it could make even *more* profit by sewing pants. *Accounting* profit considers only the difference between the revenues that a firm receives and its explicit costs. Before calculating *economic* profit, implicit or opportunity costs need to be added to the explicit costs. The implicit cost of using sewing machines to make shirts is the profit given up by not choosing the next best alternative, in this case sewing pants. These implicit or opportunity costs become important when the firm makes the "wrong" allocation (i.e., when it could have made more profit doing something else). When implicit costs are taken into consideration, accounting profits often turn into economic losses.

There are implicit costs associated with the firm's use of retained earnings or profits. It may seem logical to use funds to purchase additional capital equipment, but the firm needs to consider how else those funds could be used. Similarly, if an individual decides to use her time to start a consulting business, she may be very pleased if, at the end of a year, she has covered all her expenses and made a tidy profit. However, this individual needs to also consider what else she could have accomplished with that year of work. Perhaps she could have earned much more working for someone else. This is not to suggest that making the most money lies at the heart of all economic allocation decisions. But it certainly emphasizes that considering the alternative uses of resources and measuring the value of forgone opportunities are critical to making correct production and allocation of resources decisions.

*Social and Private Costs*

One final cost consideration that needs to be discussed before moving to the long run is the distinction between private and social costs. Social costs include the private costs that a firm incurs plus any external costs that the firm imposes on the rest of society but typically does not pay for. If a firm dumps its dirty water into a river, it may inadvertently kill some the fish. Downstream fishermen may now have to work longer and harder to catch fresh fish. Economists as well as environmentalists have long argued that firms should be accountable for all the costs associated with their production decisions. Forcing firms to take external costs into consideration increases their marginal cost. As a result, firms usually produce less and attempt to increase the price of their product. Pollution and other negative externalities have become an important area of economic analysis (refer to Chapter 22 [this volume]).

## The Long Run

In the long run, everything is variable. There are no fixed inputs, and there are no fixed costs. Decisions to enter or exit from an industry are acted on in the long run. Keep in the mind that the long run is defined as that period in which everything that the firm does can be changed. There is no specific amount of time that separates the long run from the short run; it is an industry- or even firm-specific period of time.

*Two Variable Inputs*

Conceptually, the biggest change to understand is what happens when the firm moves from one to two variable inputs. Most textbooks present the two-variable input case and then make the logical argument that proceeding to three or more variables is straightforward. The expansion of inputs to three or more is relatively easy mathematically, but output and cost diagrams become untenable when attempting to move beyond a three-dimensional diagram.

Consider the case where a firm has two variable inputs, capital and labor. Assume that the firm has determined the optimal level of output to produce, based on the demand for its product, and now needs to determine the least-cost way of producing that level of output. The input combinations and cost options that the firm faces are illustrated in Figure 10.2, panel (c). The Greek prefix *iso* means equal or same, and the isoquant curve illustrates all the combinations of capital ($K$) and labor ($L$) that produce the same quantity of output. The isocost line shows all the combinations of $K$ and $L$ that the firm could purchase or hire and spend the same amount of money.

The shape of the isoquant suggests that the firm could produce a given amount of output using a great deal of $K$ and a little $L$ or vice versa. Note that the trade-off of $K$ for $L$ is not constant, and therefore the slope of the isoquant is nonlinear. The isoquant also seems to get closer and closer to the axes but never quite touches it. Economically, this is very attractive as it suggests that the firm needs at least a small amount of each input to produce output. Machines need someone to run them, and workers require capital equipment to do work. The shape of this isoquant is based

on a famous production function called the Cobb-Douglas production function, which is discussed further in the empirical section of this chapter. In its most basic form, the Cobb-Douglas production function suggests that the output of a firm can be estimated by multiplying together the utilization of its inputs. Assume that this firm is interested in producing 200 units of output. All the *K–L* coordinates along the isoquant multiply to 200. For example, consider the following combinations of inputs: 20*K* and 10*L,* 10*K* and 20*L,* and 5*K* and 40*L.* Indeed, all the capital–labor points along this isoquant multiply to 200.

The least-cost way of producing 200 units of output will be determined by the cost of the inputs. The isocost line in the diagram illustrates all the combinations of *K* and *L* that the firm could acquire for a fixed budget, say $600. (The number 600 is chosen for illustrative purposes and yields cost-of-input prices that are intuitively appealing, but nearly any budget number could be used.) If the firm allocated all of its budget to labor, it could hire 40 labor hours or work. Hence, the hourly wage rate (*w*) is $15. Similarly, if the firm spent its entire budget on capital, it could use 20 hours of machine time. This implies that the rental rate (*r*) for capital equipment is $30 per hour. It should be noted that even if the firm owns its own capital equipment, the rental rate still applies because there is an implicit cost of using the machines for one production process as opposed to another.

This analysis assumes the firm takes the wage rate and price of capital as given and, as stated earlier, that both labor and capital are homogeneous. The slope of the isocost line is the negative of the two input price ratios, or the price of labor over the price of capital. That is, the slope equals $-w/r$, in this case −15/30, or −0.5. An understanding of the isocost line helps one understand why different firms use different combinations of inputs, even when they have the same technology and face the same production possibilities. As the wage rate falls, the isocost line becomes flatter, and the firm substitutes labor for capital. It may appear that some firms overuse labor relative to capital when in fact those firms may simply be located in areas with relatively low wage rates.

### *Economies of Scale*

Economies of scale measure how a firm's output changes when it is able to change its entire scale of operations. By definition, the discussion as well as estimates of returns to scale can take place only in the long run because the firm must be able to vary all of its inputs. There are three scale effects to consider: increasing, decreasing, and constant returns to scale.

Increasing returns to scale (RTS) means that as a firm increases the utilization of its inputs, output naturally increases but at an even faster rate. For example, if a firm doubles the utilization of its inputs, output more than doubles. Adam Smith (1776/2008) was probably the first to write about increasing RTS. In his famous example of a pin factory, Smith describes how workers specializing in the various steps of manufacturing can greatly increase the output of the firm. In addition to labor specialization, firms can also benefit by using larger, more efficient capital equipment. For many industries, this means that to produce at lower per unit cost, the scale of operation must be very large. Companies may also realize increasing RTS by receiving bulk-order discounts on the purchase of large amounts of inputs. Incorporating new technologies that become cost-effective with large-scale operations can be another source of increasing RTS. Henry Ford's development of assembly line techniques in the automobile industry dramatically lowered the per unit cost, and thereby the price, of cars.

Eventually, however, the firm may become too large to manage effectively. The scale of operations may become too complex, and communication may break down as the chain of command becomes longer and longer. Furthermore, workers may feel more and more alienated from management and the decision-making process. When this occurs, the firm experiences decreasing RTS. In this range of output, an increase in the utilization of the inputs, say by 20%, increases output, but by some smaller amount.

Constant returns to scale indicate that proportionate increases in the utilization of inputs yield a commensurate increase in the level of output. That is, changing the level of inputs, either increasing or decreasing, by 50% results in a 50% change in the level of output. Studies of returns to scale, discussed below, generally indicate that as firms grow in size, they first experience increasing RTS, then constant RTS, and finally decreasing RTS.

### *Long-Run Average Cost*

The existence of economies of scale is the primary determinant of a firm's long-run average costs. The long-run average cost curve also provides an envelope or cradle on which the firm's short-run average cost curves sit. Recall that all firms operate and produce in the short run; the long run is a planning horizon. Each of the five short-run average total curves illustrated in Figure 10.2, panel (d) represents a different-size plant or factory that a firm may decide, in the long run, to build. The inclusion of five short-run average total cost curves is for illustrative purposes. In actuality, entrepreneurs have many more long-run options to choose from.

Suppose, for example, that a firm finds itself in the short run with a demand for its product equal to output level $Q_1$. If the firm has built a factory that is too small, it may find itself on the first short-run average total cost curve ($SR - ATC_1$) and experience higher than optimal average costs. It may decide that it could appreciably lower average costs in the long run by building a larger factory and moving to $SR - ATC_2$. Building an even larger factory would lower average total costs more, but unless the firm expects a further increase in the demand for its product, it would be unwise to do so.

Notice that as the firm expands, it realizes increasing returns to scale. This is illustrated by the declining slope of

the long-run average total cost curve. Over a substantial range of output, the long-run average total cost curve is horizontal, or nearly so. This suggests that the firm's increased utilization of inputs is matched by equivalent increases in output, and therefore its (long-run) average costs are constant. This corresponds to constant returns to scale. Eventually, as the firm increases in size, it becomes too large to manage efficiently, and decreasing returns to scale or diseconomies of scale set in.

## Applications and Empirical Evidence

### Production Functions

One of the most widely tested production functions is the Cobb-Douglas (Cobb & Douglas, 1928) production function. Several factors account for Cobb-Douglas's longevity. It is relatively easy to collect the data necessary to estimate the production function. The estimation procedure itself is relatively straightforward. The economic characteristics of the production function match up reasonably well with the theoretical model of production. And perhaps most important, the empirical estimates obtained from the model have done a reasonably good job of explaining real-world production behavior.

In the two input case, the Cobb-Douglas production function can be written as $Q = A\,K^{\alpha}L^{\beta}$ where $Q$ is output, $K$ and $L$ are the capital and labor inputs, and $A$, $\alpha$, and $\beta$ are the parameters that are estimated. If there are more than two inputs, they simply get multiplied on to the equation in similar format. The summation of the estimated exponents on the inputs, in this case $\alpha + \beta$, yields an estimate of returns to scale for the industry being studied. If the sum is greater than 1, it suggests the industry is experiencing increasing returns to scale. Sums of less than 1 imply decreasing returns to scale. And if the sum of the exponents is equal to 1, the industry has constant returns scale. Furthermore, as long as each individual exponent is less than 1, the marginal product of each associated input will be declining and thereby in sync with the law of diminishing (marginal) returns.

Mansfield (1997) summarizes the Cobb-Douglas estimation results from six different studies, on 18 different industries, in five different countries. The industries studied varied widely and included examples from agriculture, manufacturing, and transportation. An examination of the estimated exponents suggests that 9 of the industries were experiencing increasing returns to scale, 6 were in the constant returns to scale range, and 3 had decreasing returns to scale. Overall, the summed exponents did not vary widely and ranged from 0.87 to 1.29. Each of the 38 estimated exponents was less than 1, supporting the law of diminishing returns. Using the estimated coefficients from the French gas industry, Mansfield was able to construct isoquants that looked very similar to those described in the theoretical model.

As appealing as these results are, the Cobb-Douglas model does have some shortcomings. Because the production function is multiplicative, it assumes that every input is used continuously. If a single input is not used, Cobb-Douglas suggests that output falls to zero. Despite the nicely shaped isoquants, the Cobb-Douglas production function does not yield the smooth bell-shaped marginal and average product curves described earlier.

These and other shortcomings have led economists to develop more realistic and hence more complicated production function specifications. One other type of production function, of which Cobb-Douglas is a special case, is the constant elasticity of substitution (CES) production function. The interested reader is referred to a classic article by Arrow, Chenery, Minhas, and Solow (1961) to study this case.

### Cost Curves

Economists have had a difficult time obtaining accurate cost data. In the first place, accounting data never include the implicit costs that are so important in economic decision making. Second, firm and industry data, when they are available, are collected on a fiscal year basis. Rarely does this coincide with either the short- or long-run time periods of economic analysis. Third, inputs are often used in different ways to produce different products. This makes it difficult to assign costs specifically to the production of any one good or service.

Nevertheless, costs, like production functions, have been studied widely. The results (Mansfield, 1997), while even more widely dispersed than production function estimates, are very interesting. Studies of long-run average cost, across a wide range of industries, over a long period of time, show little evidence of decreasing returns to scale. Rather than suggesting that decreasing returns rarely occur, it is more likely that when they become evident, entrepreneurs act quickly to eliminate them by reducing the scale of operations and perhaps creating separate companies.

Several studies, across a wide range of industries, indicate that long-run average costs may be "L"-shaped but with the horizontal portion of the L extended out to the right. That is, the industry first experiences increasing returns to scale, and then as it grows enjoys constant long-run average costs (and constant returns to scale) over a wide range of output.

A fewer number of studies suggest that increasing returns to scale are evident. This implies that the long-run average cost curve is negatively sloped. If an industry is experiencing increasing returns to scale, there is a strong incentive (as long as there is sufficient demand for the product) to increase the scale of operations because revenues are increasing and average costs are falling. Assuming demand is strong, economic theory suggests that industries will expand through the range of output in which they experience increasing returns. This is not because this is an undesirable position to be in. It is exactly the opposite: If there are increasing returns to scale, there is an incentive to grow bigger and benefit from even more increasing returns.

Clearly, the area of cost estimation is one where more careful work by economists is needed. Replicating studies with new sources of data can be a way to lend greater support (or doubt) to earlier work. The development of new economic software models and estimation procedures will help in this analysis.

## Conclusion

Whether a firm is in business to make profits or promote the common good, the importance of carefully analyzing cost and production decisions cannot be overstated. Without a firm handle on the way inputs can be combined to produce products and the cost of acquiring those inputs, a firm will surely fail.

It is hoped that this discussion, while being primarily theoretical, helps clarify some very real-world concerns. For example, the conceptual distinction between the short and long run helps to clarify the difference between shutting down a firm and exiting from the industry. Holding some inputs constant may not seem entirely logical but drives home the importance of determining just how productive individual inputs are.

Economic models are often evaluated on two fronts—how well they explain concepts and how well they predict. This analysis has focused primarily on explaining how firms make decisions. While more work needs to be done on acquiring reliable data, the correlation between theoretical explanation and reality seems strong.

## References and Further Readings


Arrow, K., Chenery, H., Minhas, B., & Solow, R. (1961). Capital-labor substitution and economic efficiency. *Review of Economics and Statistics, 43,* 225–250.

Binger, B. R., & Hoffman, E. (1998). *Microeconomics with calculus* (2nd ed.). Reading, MA: Addison-Wesley.

Cobb, C. W., & Douglas, P. H. (1928). A theory of production. *American Economic Review, 18*(Suppl.), 139–165.

Douglas, P. H. (1948). Are there laws of production? *American Economic Review, 38,* 1–48.

Ferguson, C. E. (1969). *The neoclassical theory of production and distribution.* Cambridge, UK: Cambridge University Press.

Gould, J. P., & Ferguson, C. E. (1980). *Microeconomic theory* (5th ed.). Homewood, IL: Irwin.

Hall, R. E., & Lieberman, M. (2005). *Microeconomics: Principles and applications* (3rd ed.). Mason, OH: Thompson-Southwestern.

Hirshleifer, J., Glazer, A., & Hirshleifer, D. (2005). *Price theory and applications: Decisions, markets, and information* (7th ed.). Cambridge, UK: Cambridge University Press.

Krugman, P., & Wells, R. (2009). *Microeconomics* (2nd ed.). New York: Worth.

Mahanty, A. K. (1980). *Intermediate microeconomics with applications.* New York: Academic Press.

Mansfield, M. (1997). *Applied microeconomics* (2nd ed.). New York: W. W. Norton.

Mansfield, M., & Yohe, G. (2003). *Microeconomics: Theory and applications* (11th ed.). New York: W. W. Norton.

McConnell, C. R., & Brue, S. L. (2005). *Microeconomics* (16th ed.). New York: McGraw-Hill/Irwin.

Pfouts, R. W. (1972). *Elementary economics: A mathematical approach.* New York: John Wiley.

Salvatore, D. (1991). *Schaum's outline of microeconomics* (3rd ed.). New York: McGraw-Hill.

Samuelson, P. A., & Nordhaus, W. D. (2006). *Economics* (18th ed.). New York: McGraw-Hill.

Sher, W., & Pinola, R. (1981). *Microeconomic theory: A synthesis of classical theory and the modern approach.* New York: Elsevier North-Holland.

Smith, A. (2008). *An inquiry into the nature and causes of the wealth of nations.* Hamburg, Germany: Management Laboratory Press. (Original work published 1776)

Varian, H. R. (2005). *Intermediate microeconomics: A modern approach* (7th ed.). New York: W. W. Norton.

# 11

# Profit Maximization

Michael E. Bradley
*University of Maryland, Baltimore County*

Mainstream microeconomics generally assumes that firms seek to maximize economic profit, the difference between total revenue and total economic costs. This, like many others in economics, is an unrealistically simple picture of the way that firms actually make decisions. Economists make many such unrealistic assumptions that abstract from and simplify the "real world" to explain economic phenomena and generate empirically testable predictions.

Assuming that firms act rationally to maximize profit is critical for analyzing and explaining the firm's choices of outputs of goods and factor inputs. The theory of production and cost collapses without assuming profit maximization or some other maximizing assumption. For example, a production function determines a single output from each bundle of factor inputs, but technology and resource quality define only the maximum output that can be produced with a given bundle of inputs, and there is nothing to prevent firms from producing less than the maximum output. If we assume that the firm seeks to maximize profit, it follows that it will produce only the maximum output from any given bundle of inputs or that it will use the combination of inputs that produces a given output at the lowest cost.

## Profit Maximization: The General Case

### Assumptions

Because profit is the difference between the revenue and cost generated by factor inputs and outputs, the profit-maximizing model assumes given production, revenue, and cost functions. These, in turn, are defined only for given resource quality, technology, product prices, and factor prices.

In addition to the assumptions that define production and cost functions, the simple profit-maximizing model assumes that individual firms have no effect on price (perfect competition), that there is a predictable relationship between output and price (monopoly), or that the rival firms in imperfectly competitive markets react predictably to each other's actions in the market. It also generally ignores or assumes away risk and uncertainty, or at least that the firm's economic expectations are unchanged.

### Revenue, Cost, and Profit

In its simplest form, the problem for the firm is to maximize profit ($\pi$), which is the difference between total revenue ($TR$) and total economic or opportunity cost ($TC$)—that is, to maximize

$$\pi = TR - TC. \tag{1}$$

The mathematical conditions for maximum profit apply to all profit-maximizing firms' choices in product and factor markets.

### Profit and Output

The profit-maximizing firm will choose the output ($q$) that maximizes

$$\pi(q) = TR(q) - TC(q), \tag{2}$$

where $\pi(q)$, $TR(q)$, and $TC(q)$ are profit, revenue, and cost functions, respectively. Total revenue is simply the price of the product ($P$) multiplied by output ($q$), or

$$TR(q) = Pq. \tag{3}$$

Output decisions are made in the short run, in which at least one of its factor inputs is fixed and does not vary with

output. The firm's short-run total cost is the sum of its fixed and variable cost, or

$$TC(q) = FC + VC\,(q) \tag{4}$$

where $TC(q)$ is the short-run total cost function, $VC(q)$ is the variable cost function, and $FC$ is fixed cost that does not vary with output. Substituting (4) and (3) into (2), the profit function becomes

$$\pi\,(q) = TR(q) - TC(q) = Pq - [FC + VC\,(q)]. \tag{5}$$

## Marginal Profit

Solving (5) for the output that maximizes profit is straightforward, *if* the firm has complete information—the price at which it can sell its output, its production function, and factor prices. However, firms seldom if ever have all of the information to find their profit-maximizing outputs by solving a profit equation. Economists in the "Austrian" tradition do not view profit maximization as solving an equation based on known information, but entrepreneurial "groping," or looking for ways to increase profit in a world of incomplete and diffused information.

Firms arrive at their maximum profit by trial and error, making small changes in output (and price if they can affect the price at which they sell) as long as these changes cause profits to rise. In economic terms, they will maximize profit at the margin where the last change in output causes no change in profit. If the firm changes its output by some small $\Delta q$, profit will change by

$$\Delta\pi(q) = \Delta TR(q) - \Delta TC(q). \tag{6}$$

If $\Delta TR(q) > \Delta TC(q)$, $\Delta\pi(q) > 0$, the firm will increase profit by increasing output. If $\Delta TR(q) < \Delta TC(q)$, $\Delta\pi(q) < 0$, the firm will increase profit by reducing output. If $\Delta TR(q) = \Delta TC(q)$ and $\Delta\pi(q) = 0$, profit is stationary—either a maximum or minimum value.

The firm's marginal profit is the rate at which profit changes with output, or the change in profit per unit of change in output. Dividing (6) through by $\Delta q$, marginal profit is

$$\frac{\Delta\pi(q)}{\Delta q} = \frac{\Delta TR(q)}{\Delta q} - \frac{\Delta TC(q)}{\Delta q} = MR(q) - MC(q), \tag{7}$$

where $MR(q)$ is the marginal revenue and $MC(q)$ is the marginal cost of the last change in output.

## Choice of Output

Although the firm does not have all of the information assumed in the production, cost, and profit functions, it will continue to make trial-and-error adjustments of output until profit stops rising at a stationary value with

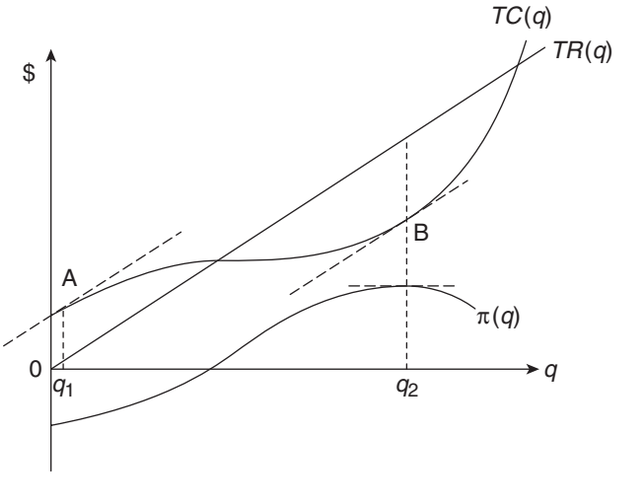


**Figure 11.1** Minimum and Maximum Profit

$$\frac{\Delta\pi(q)}{\Delta q} = MR(q) - MC(q) = 0, \text{ or} \tag{8a}$$

$$MR(q) = MC(q). \tag{8b}$$

This is the necessary condition for maximizing profit. However, it is also the necessary condition for minimizing profit.

For example, in Figure 11.1, $MR(q) = MC(q)$, and marginal profit is zero at $q_1$ and $q_2$. At $q < q_1$, $\Delta\pi(q)/\Delta q < 0$, and at $q > q_1$, $\Delta\pi(q)/\Delta q > 0$, which means that moving away from $q_1$ increases profit, and profit is at a local minimum. The opposite is true at $q_2$. At $q < q_2$, $\Delta\pi(q)/\Delta q > 0$, and at $q > q_2$, $\Delta\pi(q)/\Delta q < 0$, which means that moving away from $q_2$ will cause profit to fall, and profit is at a local maximum.

A firm is maximizing profit if it satisfies the necessary condition,

$$\frac{\Delta\pi(q)}{\Delta q} = 0, \text{ or } MR(q) = MC(q), \tag{9}$$

and the sufficient condition,

$$MR(q) > MC(q) \text{ and } \frac{\Delta\pi(q)}{\Delta q} > 0 \tag{10a}$$

for smaller outputs, and

$$MR(q) < MC(q) \text{ and } \frac{\Delta\pi(q)}{\Delta q} < 0 \tag{10b}$$

for larger outputs.

The sufficient condition for maximum profit is satisfied if the slope of $MR(q)$ is less than the slope of $MC(q)$—that is, if

$$\frac{\Delta MR(q)}{\Delta q} < \frac{\Delta MC(q)}{\Delta q}, \tag{11}$$

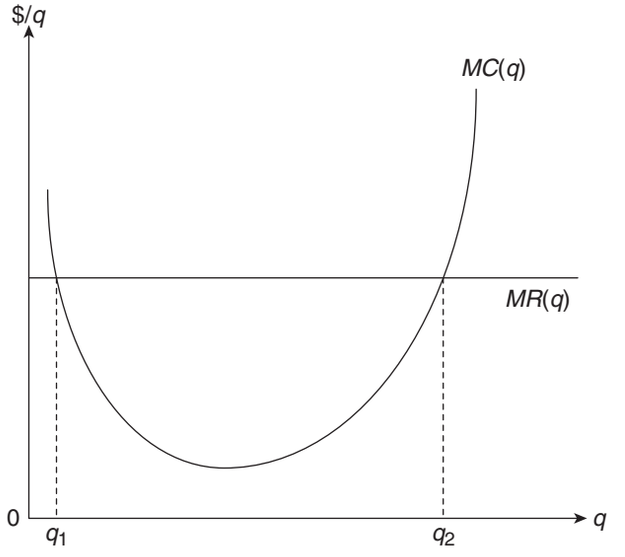


**Figure 11.2** Marginal Cost, Marginal Revenue, and Minimum and Maximum Profit

and $MC(q)$ cuts $MR(q)$ "from below."

In Figure 11.2, for example, $MR(q)$ is constant, and $MR(q) = MC(q)$ at $q_1$ and $q_2$. At $q_1$, $\Delta MR(q)/\Delta q = 0 > \Delta MC(q)/\Delta q < 0$, and profit is at a local minimum. At $q_2$, $\Delta MR(q)/\Delta q = 0 < \Delta MC(q)/\Delta q > 0$, which satisfies the sufficient condition for maximum profit.

## Shutting Down

The firm's profit-maximizing (or loss-minimizing) choice may be to "shut down" and produce zero output. A profit-maximizing firm will produce its profit-maximizing output $\overline{q} > 0$ only if

$$\pi(\overline{q}) \geq \pi(0). \tag{12}$$

If $\pi(\overline{q}) = \pi(0)$, the firm would do no better by producing than shutting down, and it would be at its shutdown point. If $\pi(\overline{q}) < \pi(0)$, the firm would do better by shutting down and producing $q = 0$. This would be the case if the firm's total revenue from producing, $TR(\overline{q})$, is less than the *variable cost* of producing $\overline{q}$. In the long run, with all factor inputs variable, all long-run cost, $LRTC(q)$, is variable, $\pi(0) = 0$, and the firm will shut down in the long run if $TR(\overline{q}) < LRTC(\overline{q})$ and $\pi(\overline{q}) < 0$.

In the short run, the firm incurs the costs of its fixed factor inputs ($FC$), even if the firm does not produce, so short-run profit at zero output is

$$\begin{aligned} \pi(0) &= TR(0) - TC(0) = \\ TR(0) &- VC(0) - FC = -FC \end{aligned} \tag{13}$$

because $TR(0) = VC(0) = 0$. The firm will produce its profit-maximizing output $\overline{q} > 0$ in the short run if $\pi(\overline{q}) \geq \pi(0) = -FC$.

The firm is at its short-run shutdown point if $\pi(\overline{q}) = \pi(0) = -FC$. At this point,

$$\pi(\overline{q}) = TR(\overline{q}) - [FC + VC(\overline{q})] = -FC \tag{14a}$$

$$\pi(\overline{q}) = TR(\overline{q}) - VC(\overline{q}) = 0 \tag{14b}$$

$$TR(\overline{q}) = VC(\overline{q}). \tag{14c}$$

Dividing (10a) through by $\overline{q}$, the firm is at its short-run shutdown point if

$$\frac{P\overline{q}}{\overline{q}} = \frac{VC(\overline{q})}{\overline{q}} \tag{14d}$$

$$P = AVC(\overline{q}). \tag{14e}$$

If the firm is exactly at the shutdown point, we do not know whether it would produce or shut down because there is no net benefit from producing. If $P > AVC(\overline{q})$ and $TR(\overline{q}) > VC(\overline{q})$, it will produce in the short run, even if its revenue does not cover all of its costs. If $P < AVC(\overline{q})$ and $TR(\overline{q}) < VC(\overline{q})$, the firm does better by shutting down and producing no output.

The only difference between the short-run and long-run shutdown points is the absence of fixed costs in the long run because all of the firm's inputs and costs are variable. In the long run, then, the firm is at the shutdown point if

$$\pi(\overline{q}) = TR(\overline{q}) - LRTC(\overline{q}) = 0 \tag{15a}$$

$$TR(\overline{q}) = LRTC(\overline{q}). \tag{15b}$$

Dividing (15a) through by $\overline{q}$, the firm is at its long-run shutdown point if

$$P = LRAC(\overline{q}). \tag{15c}$$

In short, to produce its profit-maximizing output in the long run, the firm's revenue must at least cover its long-run total costs, all of which are variable, or the price of its output must at least cover its long-run average cost. Again, the firm realizes no net benefit from producing at the shutdown point.

## Producer Surplus: Net Benefit From Producing

If a firm can realize more profit or incur a smaller loss by producing than by shutting and producing zero output, it will realize a net benefit from producing, or producer surplus. Because the firm would realize no net benefit from producing if $TR(q) = VC(q)$, we can define producer surplus as

$$PS(q) = TR(q) - VC(q) \tag{16a}$$

in the short run, and

$$PS(q) = TR(q) - LRTC(q) \tag{16b}$$

in the long run with all costs variable.

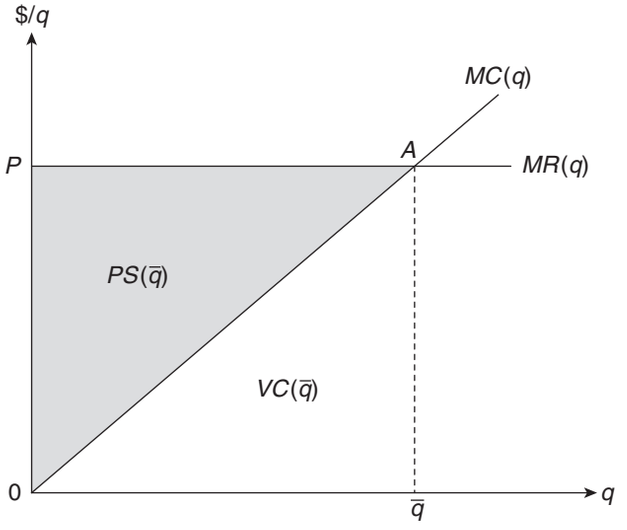


**Figure 11.3** Maximum Profit and Producer Surplus

Producer surplus is illustrated graphically in Figure 11.3. For simplicity, assume a linear marginal cost function, $MC(q)$. If the firm is a price taker and faces a constant price for its output, then price and marginal revenue are identical, $P \equiv MR(q)$ The firm maximizes profit at $\overline{q}$, where $P \equiv MR(\overline{q}) = MC(\overline{q})$ The firm's total revenue is $TR(\overline{q}) = P\overline{q} = AREA\ OPA\overline{q}$. Variable cost is the sum of the marginal cost of outputs between 0 and $\overline{q}$,

$$VC(q) = \sum_{q\ 0}^{\overline{q}} MC(q), \tag{17}$$

or the area under the marginal cost curve $0A\ \overline{q}$. Producer surplus is

$$\begin{gathered} PS(\overline{q}) = TR(\overline{q}) - VC(\overline{q}) = \\ AREA\ OPA\overline{q} - AREA\ OA\overline{q} = \\ AREA\ OPA. \end{gathered} \tag{18}$$

In the short run, the firm may realize producer surplus, even if it incurs an economic loss, as long as it covers its variable costs. However, in the long run, there is no distinction between producer surplus and profit because long-run costs are all variable—that is, in the long run,

$$PS(q) \equiv \pi(q) = TR(q) - LRTC(q). \tag{19}$$

A firm realizes a rent if it is more than compensated for producing rather than not producing. The firm in Figure 11.3, for example, earns no rent on the last, or marginal, unit produced. However, it does realize an intramarginal rent on the units up to the marginal unit because it receives a price and marginal revenue above the marginal cost of these units, and $TR(\overline{q}) > VC(\overline{q})$. Because variable cost just compensates the firm for producing rather than not producing, producer surplus is a form of intramarginal rent.

## Choice of Factor Inputs

If the firm produces its profit-maximizing output, obviously it must employ its profit-maximizing factor inputs. We can simplify our analysis by assuming that there are only two factors of production, labor ($n$) and capital ($k$). Combinations of factors will generate profit of

$$\pi(n,k) = TR(n,k) - TC(n,k), \tag{20}$$

assuming a production function $q = q(n,k)$ and a profit function of

$$\pi(n,k) = Pq(n,k) - TC(n,k). \tag{21}$$

With a variable labor input and fixed capital input in the short run,

$$\begin{gathered} \pi(n,k) = Pq(n,\hat{k}) - TC(n,\hat{k}) = \\ Pq(n,\hat{k}) - wn(q) - P_k\hat{k}, \end{gathered} \tag{22}$$

where $w$ is the wage per unit of labor and $P_k\hat{k}$ is the cost of the fixed input, or $FC$.

With capital fixed, we can express output as a function of the variable labor input alone because $\hat{k}$ is a parameter. This gives us the short-run profit function

$$\begin{gathered} \pi(n,k) = Pq(n) - wn(q) - P_k\hat{k} = \\ TR(n) - VC(n) - FC. \end{gathered} \tag{23}$$

The firm reaches its profit-maximizing input of labor by making trial-and-error adjustments as long as they cause profit to rise. Changing the variable input, $n$ in this case, affects profit by changing output and revenue and by changing cost. From (23), changing $n$ by some small $\Delta n$ changes profit by

$$\Delta\pi(n) = \Delta TR(n) - \Delta TC(n). \tag{24}$$

The marginal profit is the rate at which $\pi(n)$ changes with $n$, or the change in profit per unit of change in labor,

$$\begin{gathered} \frac{\Delta\pi(n)}{\Delta n} = \frac{\Delta TR(n)}{\Delta n} - \frac{\Delta TC(n)}{\Delta n} = \\ MRP(n) - MFC(n), \end{gathered} \tag{25a}$$

where

$$MRP(n) = \Delta TR(n)/\Delta n \tag{25b}$$

is the marginal revenue product of the variable factor input ($n$), and

$$MFC(n) = \Delta TC(n)/\Delta n \equiv \Delta VC(n)/\Delta n \tag{25c}$$

is the marginal factor cost of the variable input ($n$).

If $MRP(n) > MFC(n)$, profit varies directly with the labor input, and the firm would increase its profit by increasing its labor input. Conversely, if $MRP(n) < MFC(n)$, profit varies inversely with the labor input, and the firm would reduce its labor input to increase profit. The firm will stop changing inputs when profit stops rising. This occurs where the marginal profit from the last change is

$$\frac{\Delta\pi(n)}{\Delta n} = MRP(\overline{n}) - MFC(\overline{n}) = 0 \tag{26a}$$

$$MRP(\overline{n}) = MFC(\overline{n}), \tag{26b}$$

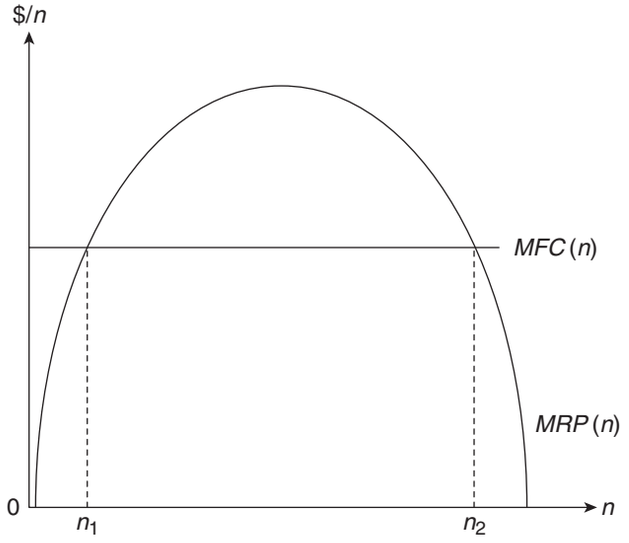


**Figure 11.4** Profit-Maximizing Labor Input

which is the necessary condition for a profit-maximizing labor input ($\bar{n}$). This is true at the minimum profit as well as the maximum, so we have to specify that $\pi(n)$ is at its maximum if $MRP(\bar{n}) = MFC(\bar{n})$, and $MRP(\bar{n}) > MFC(\bar{n})$ for smaller inputs and $MRP(\bar{n}) < MFC(\bar{n})$ for larger inputs.

In Figure 11.4, for example, $MFC(q)$ is constant, and $MRP(n) = MFC(n)$ at $n_1$ and $n_2$. Both of these inputs satisfy the necessary condition for a maximum (26a & b). However, the sufficient condition is satisfied only at $n_2$, which is the firm's profit-maximizing labor input. Profit is at its minimum value at $n_1$. Graphically, the necessary and sufficient conditions for maximum $\pi(n)$ are satisfied where $MRP(n) = MFC(n)$, and $MFC(n)$ "cuts $MRP(n)$ from below." At $n_1$, $MRP(n) = MFC(n)$, $MFC(n)$ "cuts $MRP(n)$ from above," and $\pi(n)$ is at a minimum.

## Profit-Maximizing Factor Inputs and Output

We can verify easily that if the firm is producing its profit-maximizing output where $MR(\bar{q}) = MC(\bar{q})$, it must employ its profit-maximizing factor inputs where $MFC(\bar{n}) = MRP(\bar{n})$. First rewrite the equations for $MRP(n)$ and $MC(q)$.

$$MRP(n) = \frac{\Delta TR(n)}{\Delta n} = \frac{\Delta q}{\Delta n}\frac{\Delta TR}{\Delta q} = MP(n)MR(q) \quad (27a)$$

$$MC(q) = \frac{\Delta TC(q)}{\Delta q} = \frac{\frac{\Delta TC}{\Delta n}}{\frac{\Delta q}{\Delta n}} = \frac{MFC(n)}{MP(n)}. \quad (27b)$$

If the firm employs its profit-maximizing labor input, it will produce its profit-maximizing output because

$$MFC(\bar{n}) = MP(\bar{n})MR(\bar{q}) \quad (28a)$$

$$\frac{MFC(\bar{n})}{MP(\bar{n})} = MR(\bar{q}) \quad (28b)$$

$$MC(\bar{q}) = MR(\bar{q}). \quad (28c)$$

Similarly, if the firm is producing its profit-maximizing output, it must employ the profit-maximizing factor input, because from (27b),

$$MC(\bar{q}) \equiv \frac{\Delta TC}{\Delta q} = \frac{\frac{\Delta TC}{\Delta n}}{\frac{\Delta q}{\Delta n}} = \frac{MFC(\bar{n})}{MP(\bar{n})} = MR(\bar{q}) \quad (29a)$$

$$MFC(\bar{n}) = MR(\bar{q})MP(\bar{n}) = MRP(\bar{n}). \quad (29b)$$

Thus, profit maximization is not a single choice but a choice of the profit-maximizing output and the profit-maximizing factor inputs.

## Average Revenue Product and the Shutdown Point

The firm's average revenue per unit of a variable input (labor, in the above analysis) is the average revenue product,

$$ARP(n) = \frac{TR(n)}{n}, \quad (30a)$$

where $n$ is the labor input and $TR(n)$ is the total revenue generated by labor. The revenue per unit of labor depends on the output per unit of labor, or average product,

$$AP(n) = \frac{q(n)}{n}, \quad (30b)$$

and the average revenue per unit of output,

$$AR(q) = \frac{TR(q)}{q}. \quad (30c)$$

Combining (30a), (30b), and (30c), average revenue product is

$$ARP(n) = \frac{TR(n)}{n} = \frac{TR(q)}{q}\frac{q}{n} = AR(q)AP(n). \quad (30d)$$

If the firm sells all units of its output at the same price ($P$), then $AR(q) \equiv P$, and average revenue product is

$$ARP(n) = P \bullet AP(n). \quad (30e)$$

We know that the firm will shut down and produce $q = 0$ if its maximum profit from producing, $\pi(q)$, is less than the profit from not producing at all, $\pi(0)$. This means that the firm will hire its profit-maximizing labor input, $\bar{n} > 0$, only if its profit is no lower than if it shut down and hired no labor input—that is, it will produce its profit-maximizing output and hire its profit-maximizing labor input, $\bar{n} = n(\bar{q})$, only if $\pi(\bar{q}) \geq \pi(0)$ and employ its profit-maximizing labor input only if $\pi[n(\bar{q})] = \pi(\bar{n}) \geq \pi(0)$.

At the shutdown point, $\pi(\bar{n}) = \pi(0)$. In the short run,

$$\begin{aligned} \pi(0) &= TR(0) - TC(0) \\ &= TR(0) - VC(0) - FC = -FC, \end{aligned} \quad (31)$$

so the firm is at its shutdown point in the labor market if

$$\pi(\bar{n}) = TR(\bar{n}) - VC(\bar{n}) - FC = -FC \tag{32a}$$

$$TR(\bar{n}) = VC(\bar{n}) = w\bar{n} \tag{32b}$$

$$\frac{TR(\bar{n})}{\bar{n}} = \frac{w\bar{n}}{\bar{n}} \tag{32c}$$

$$ARP(\bar{n}) = w. \tag{32d}$$

If the firm has to pay a wage above $ARP(\bar{n})$, it will shut down because its revenue will be insufficient to cover its variable costs.

## Choosing the Profit-Maximizing Capital Input

Usually, the analysis of short-run profit maximization assumes a fixed capital input and variable input, but this does not have to be the case. We could also assume a fixed labor input and variable capital input, in which case the conditions for choosing the profit-maximizing capital input would be the same as for the profit-maximizing labor input—namely,

$$\frac{\Delta\pi(\bar{k})}{\Delta k} = MRP(\bar{k}) - MFC(\bar{k}) = 0 \tag{33a}$$

$$MRP(\bar{k}) = MFC(\bar{k}). \tag{33b}$$

However, capital is a much broader and more complex class of inputs than labor. It is easy to imagine a unit of labor as a worker-hour, worker-day, and so on. It is not so easy to imagine a unit of capital because capital includes financial capital (loanable funds), human capital, machinery and other producer durables, and structures. Moreover, current outlays on capital generate revenue and profit (if any) in the future, and the future is uncertain. This means that we have to introduce some special features of capital that affect both its *MRP* and its price and *MFC*.

First, there is the issue of time, which Alfred Marshall (1961) termed "the source of many of the greatest difficulties in economics" (p. 109). Because we do not know the future exactly, the revenue from capital is an expected value—that is,

$$TR(k) = expected\ TR(k) \tag{34a}$$

$$MRP(k) = \frac{expected\ \Delta TR(k)}{\Delta k}. \tag{34b}$$

Time also complicates things because, even with no uncertainty, risk, or inflation, \$1 today is worth more than \$1 tomorrow, next year, or 10 years from now. The present value of \$1 in the future is \$1 discounted by a pure real interest rate ($r$) after $t$ periods in the future is

$$PV(1) = \frac{1}{(1+r)^t}. \tag{35}$$

This means that $MRP(k)$ is the present value of the expected future revenue, or

$$MRP(k) = \frac{expected\ \Delta TR(k)}{\Delta k}\left(\frac{1}{(1+r)^t}\right). \tag{36}$$

It also means that $MRP(k)$ can change if economic expectations and/or the time preference for income now over income in the future change.

Uncertainty and risk also affect the price or user cost of capital, $P_k$. If firms finance their capital outlays by borrowing, they have to pay interest to lenders. If they self-finance their capital outlays, they forgo the interest they could have earned on the funds. Risk and uncertainty are real costs of employing capital, which means that the cost of capital includes a risk premium, $R$. The risk premium is zero for completely safe capital projects and varies directly with the riskiness of projects. Specifically, $R$ is a function of the probability of the success from the project ($p$) and the variance of returns—algebraically, $R_i = R_i(p, \sigma)$, where $p$ is the probability of success and $\sigma$ is the variance. The degree of risk and $R_i$ vary inversely with $p$ and directly with $\sigma$.

Finally, we have to consider that capital depreciates with use and that the depreciation rate, $d$, is another element of the cost or price of using capital. If we put all of the elements together, the price of a \$1 capital outlay becomes

$$P_k = 1 + r + R(p, \sigma) + d. \tag{37}$$

We can apply the necessary and sufficient conditions and notations for choosing the profit-maximizing input of variable labor to choosing the profit-maximizing variable capital input. However, we have to consider the greater complexity of the capital decisions by the firm.

## Profit Maximization in the Long Run

The analysis of profit maximization in the long run is quite straightforward—in some respects, a simpler problem than short-run profit maximization.

With labor and capital both variable in the long run, the firm will minimize the cost of any output where the additional output per dollar spent is equal for the last units of labor ($n$) and capital ($k$) employed, or where

$$\frac{MP(n)}{w} = \frac{MP(k)}{P_k}, \tag{38a}$$

from which it is obvious that

$$\frac{w}{MP(n)} = \frac{P_k}{MP(k)} = MC(q) = LRMC(q). \tag{38b}$$

Since $LRTC(q)$ is the minimum cost of producing any output, (38a) is true at all points on the $LRTC(q)$ curve, and (38b) is true at all points on the $LRMC(q)$ curve.

The firm maximizes profit in the long run at outputs and factor inputs of $\bar{q}$, $\bar{n}$, and $\bar{k}$, where

$$\frac{P_k}{MP(\bar{k})} = \frac{w}{MP(\bar{n})} = MC(\bar{q}) = LRMC(\bar{q}) = MR(\bar{q}) \tag{39a}$$

$$\frac{w}{MP(\bar{n})} = MR(\bar{q}) \tag{39b}$$

$$w = MR(\overline{q})MP(\overline{n}) = MRP(\overline{n}) \tag{39c}$$

$$\frac{P_k}{MP(\overline{k})} = MR(\overline{q}) \tag{39d}$$

$$P_k = MR(\overline{q})MP(\overline{k}) = MRP(\overline{k}). \tag{39e}$$

This satisfies the necessary conditions for profit-maximizing output and factor inputs. As in the short run, if the firm is producing its profit-maximizing output, it must be employing the profit-maximizing inputs of labor and capital.

As in the short run, the shutdown point in the long run is reached if the total revenue from the firm's profit-maximizing output just equals variable cost. However, with all inputs variable, all long-run costs are variable. Thus, the firm reaches its shutdown point if, at its profit-maximizing output, $\overline{q}$,

$$TR(\overline{q}) = LRTC(\overline{q}) \tag{40a}$$

$$P = LRAC(\overline{q}). \tag{40b}$$

With no fixed cost, there is no distinction in the long run between profit and producer surplus because long-run $\pi(0) = 0$, and producer surplus at any $q > 0$ is $PS(q) = TR(q) - LRTC(q) \equiv \pi(q)$.

The general conditions for profit maximizing apply to all profit-maximizing firms. Now, we can apply this analysis to profit-maximizing choices and their implications for economic efficiency in different type of markets.

## Profit Maximizing in the Perfectly Competitive Economy

Perfect competition is the "zero-friction" case in economics. In terms of profit-maximizing choices, the essential assumed property of perfect competition is that all individual firms are price takers in product and factor markets. The individual firm is too small relative to the market to affect the prices at which it sells its output or purchases the use of factors of production. For the perfectly competitive firm, product and factor prices are constants that do not vary with the firm's output or employment of factor inputs.

### Choice of Output

Like any profit maximizer, the perfectly competitive firm maximizes profit at the output $(\overline{q})$, where $MR(\overline{q}) = MC(\overline{q})$ and $\Delta MR(q)/\Delta q < \Delta MC(q)/\Delta q$. The firm's marginal revenue function is

$$MR(q) = P + \frac{\Delta P}{\Delta q}. \tag{41a}$$

If the firm can sell any output it could produce at a constant $P$, then $\Delta P/\Delta q \equiv 0$, and marginal revenue is identical with price because

$$MR(q) = P + \frac{\Delta P}{\Delta q} = P + 0 \equiv P. \qquad MR(q) \equiv P. \tag{41b}$$

It follows that, for a price-taking firm,

$$\frac{\Delta MR}{\Delta q} \equiv 0. \tag{41c}$$

Graphically, if $P$ is constant, $\Delta TR(q)/\Delta q = MR(q) \equiv P$, and $MR(q) \equiv P$ is simply a horizontal line at $P$, as in Figure 11.2.

The necessary condition for maximizing profit for a price-taking firm is where

$$MC(\overline{q}) = MR(\overline{q}) = P. \tag{42}$$

In Figure 11.2, $MC(q) = MR(q) \equiv P$ and $\Delta\pi(q)/\Delta q = 0$ at $q_1$ and $q_2$. Profit is at a minimum at $q_1$ because $\Delta MC(q)/\Delta q < 0 < \Delta MR/\Delta q = 0$. At $q_2$, $\Delta MC(q)/\Delta q > 0 > \Delta MR/\Delta q = 0$ , which is necessary and sufficient for maximum profit. Because the price taker's $MR$ is constant, it will maximize profit where $P$ = rising $MC(q)$, as at $q_2$ in Figure 11.2. This is not necessarily true if the firm is not a price taker.

### The Firm's Shutdown Point and Supply Curve

Supply curves or supply functions exist only in perfectly competitive markets in which the firms are price takers. Plug in a price, and the firms supply their profit-maximizing output, where $MC(q) = P$, which means that only price-taking firms have supply curves, and market or industry supply curves only exist in markets in which the firms are price takers. If a firm is large enough relative to the market that its output decisions affect the price, there is no price for the firm to "take," and consequently there is no supply curve.

Of course, if the price is below the firm's shutdown price and its profit-maximizing output cannot cover variable cost, it will not produce or supply any input.

Remember that at the shutdown point, $TR(\overline{q}) = VC(\overline{q})$, $P = AVC(\overline{q})$, and $\pi(\overline{q}) = \pi(0) = -FC$. For a price taker, $P \equiv MR(\overline{q})$, and it is at its shutdown point where

$$P \equiv MR(\overline{q}) = MC(\overline{q}) = AVC(\overline{q}) \tag{43}$$

which occurs at the minimum point on the $AVC(q)$ curve where $MC(q) = \min AVC(q)$. The price-taking firm will shut down and produce $q = 0$ if $P < \min AVC(q)$. For any $P > \min AVC(q)$, the firm will produce and supply the quantity $(q_s)$, where $P = MC(q_s) \geq AVC(q)$. This generates the firm's supply curve $q_s(P)$ in Figure 11.5.

In Figure 11.6, the perfectly competitive firm's $q_s(P)$ is the portion of $MC(q)$ above the minimum point on $AVC(q)$, where $MC(q) = \min AVC(q)$. If $MC(q)$ and $AVC(q)$ are

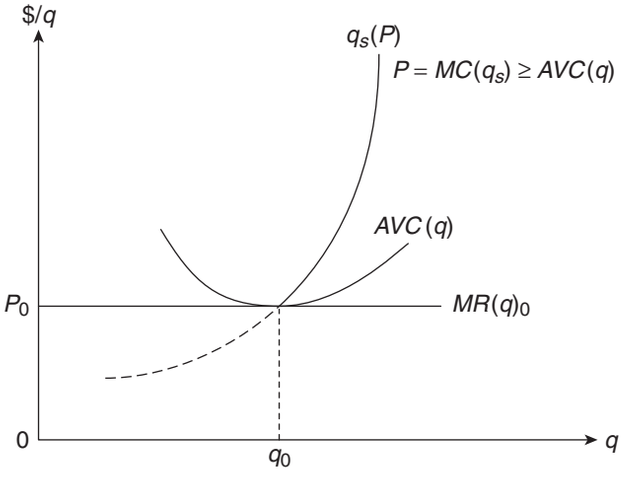


**Figure 11.5** Shutdown Point and the Competitive Firm's Supply Curve

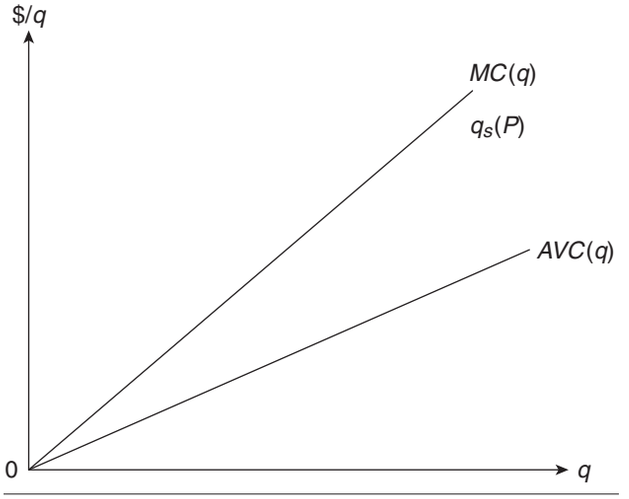


**Figure 11.6** Marginal Cost, Average Variable Cost, and Competitive Firm's Supply Curve

linear, as in Figure 11.6, $MC(q) > AVC(q)$ for all $q > 0$, and the entire $MC(q)$ curve is the firm's short-run supply curve.

## Choice of Inputs by the Perfectly Competitive Firm

In an economy in which all markets are perfectly competitive, the individual firms are price takers in product and factor markets—that is, the prices of the firm's output and of the factor inputs are constants. If labor is a variable factor input, the firm maximizes profit, $\pi(n)$, where $MFC(\overline{n}) = MRP(\overline{n})$ and $\Delta MFC(n)/\Delta n > \Delta MRP(n)/\Delta n$, which satisfies the necessary and sufficient conditions for maximum $\pi(n)$.

As a price taker in the labor market facing a constant wage ($w$), the marginal factor cost of labor is simply the wage. If $w$ is constant, then $\Delta w/\Delta n \equiv 0$ and

$$MFC(n) = w + n\frac{\Delta w}{\Delta n} \equiv w. \tag{44}$$

The firm in a perfectly competitive labor market then will maximize profit by employing labor where

$$MFC(\overline{n}) \equiv w = MRP(\overline{n}) \tag{45a}$$

and

$$\Delta MFC(n)/\Delta n \equiv \Delta w(n)/\Delta n \equiv 0 > \Delta MRP(n)/\Delta n. \tag{45b}$$

These are the necessary and sufficient conditions for the profit-maximizing labor input, $\pi(n)$. Because the firm in a competitive factor market is a price taker, this means that the firm will maximize $\pi(n)$ where $w = MRP(\overline{n})$ on the downward sloping segment of $MRP(n)$, as in Figure 11.4.

### *Marginal Revenue Product (MRP) and Value of the Marginal Product (VMP)*

The marginal revenue product of an input—labor ($n$), for example—is the value of the last unit employed to the profit-maximizing firms that employ labor. We have already shown that $MRP(n) = MP(n)MR(q)$. The value, or marginal benefit, of the last unit of labor employed to the consumers of the output it produces is the value of the marginal product of labor,

$$VMP(n) = MP(n)P, \tag{46a}$$

because the price of the output $P$ is the marginal benefit that consumers derive from it.

If the firm is a price taker in the sale of its output,

$$\Delta P/\Delta q \equiv 0, \text{ and } MR(q) = P + (\Delta P/\Delta q)q \equiv P. \tag{46b}$$

This means that for a price taker in the product market,

$$MP(n)MR(q) \equiv MP(n)P \tag{47a}$$

$$MRP(n) \equiv VMP(n). \tag{47b}$$

The value of the last unit of labor hired to the profit-maximizing firm is the same as its value to the consumers of the goods it produces.

If we combine these results, when the perfectly competitive firm employs its profit-maximizing input of labor ($n$) or any other variable factor input,

$$w = MRP(\overline{n}) = VMP(\overline{n}). \tag{48}$$

This maximizes the net benefits to the firm and the consumers of their output. *This is true only if the firm is a price taker in the product market.*

### The Shutdown Point and the Firm's Factor Demand

Demand functions, or demand curves, generate a single quantity demanded at a given price and exist only if individual buyers are price takers who cannot affect the market-determined price. In the factor market, this means that the firms will have demand curves for variable factors of production, such as labor, only if they are price takers in the factor market.

Thus, the quantity of labor demanded $(n_d)$ by a firm in a competitive labor market will be where

$$w = MRP(n_d). \tag{49}$$

If the firm is also a price taker in the product market,

$$w = MRP(n_d) = VMP(n_d). \tag{50}$$

A firm will shut down and hire $n = 0$ if it cannot cover the variable cost of the output it produces. If the firm is a price taker in the labor market, it is at its shutdown point if $w = MRP(\overline{n}) = ARP(\overline{n})$. Because $MRP(n) = ARP(n)$ only at the maximum $ARP(n)$, the firm's shutdown point is where

$$w* = MRP(n) = \max ARP(n), \tag{51}$$

where $w*$ is the "shutdown" wage.

Obviously, a profit-maximizing firm would never hire a quantity of labor or any variable input if its $MRP(n) < 0$ because it is technologically inefficient. Even if labor were free ($w = 0$), hiring labor beyond the point where $MP(n) = MRP(n) = 0$ (and $q(n)$ is at its maximum) would reduce profit.

Putting all of this together, in a competitive factor market in which the firms are price takers, the profit-maximizing choices of the firm generate its factor demand curve, which is the segment of its $MRP(n)$ between the shutdown wage and $MRP(n) = 0$.

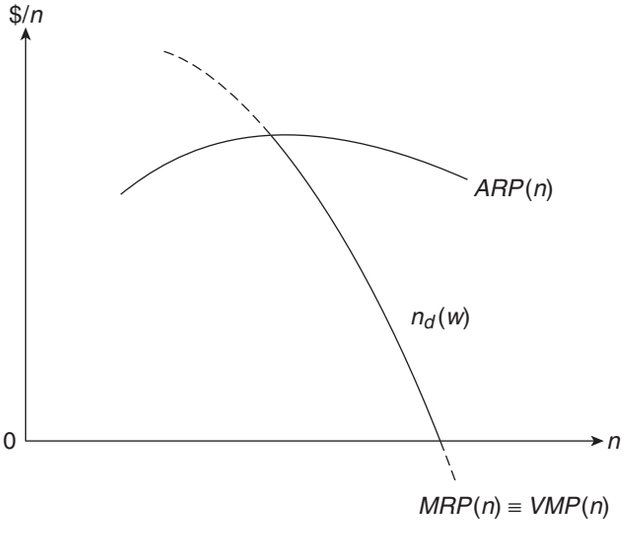


**Figure 11.7** Shutdown Point and Competitive Firm's Demand for Labor

Thus, in Figure 11.7, the firm's demand for labor is $n_d(w)$, on which $0 \leq w = MRP(n) \equiv VMP \leq \max ARP(n)$. If the firm is not a price taker in the labor market—a monopsonist, for example—it has no demand for labor because its choice of labor input determines the wage, and there is no constant wage that the firm must "take."

## Market Power and Profit Maximization

Firms with market power are not price takers. If a firm's output is large relative to the market or if it can differentiate its product from that of other firms, then its output will affect the price at which it sells. Unlike the price taker in perfect competition, a firm with market power can raise its price without losing all of its customers and reducing the quantity of its output demanded to zero, and it sell more of its output by lowering its price.

If the firm is a large user of a factor of production relative to the factor market, its input decisions affect the price of the input. The conditions for profit maximization are the same whether firms have market power or not, but the existence of market power produces some important differences from the "zero-friction" perfectly competitive case.

### Monopoly in the Product Market

The simplest case in which an individual firm has market power in the sale of its output is a monopoly, which is the only firm in the market. The monopoly model can be extended to markets that are not strictly monopolies, if the individual firm is large relative to the market and does not consider the actions of other firms.

#### *Monopoly Profit-Maximizing Output*

The necessary and sufficient conditions for maximizing profit by a monopolist are familiar: $MR(\overline{Q}) = MC(\overline{Q})$ (necessary) and $\Delta MC(Q)/\Delta Q > \Delta MR/\Delta Q$ (sufficient), where $Q$ is the quantity produced for the entire market. However, because the monopolist is not a price taker, some differences from the competitive case have important implications for the relationship between the interest of the firm in maximizing its profit and society's interest in efficiency and welfare.

#### *Price and Marginal Revenue*

A monopolist faces a negatively sloped market demand curve, so price is not constant for all outputs. It is more convenient if we think of the monopolist facing the inverse market demand curve, which expresses price as a function of quantity demanded. With a negatively sloped market demand curve, $\Delta Q_D/\Delta P < 0$ and $\Delta P/\Delta Q_D < 0$, so the price at which the monopolist sells its output is a function of output, $P(Q)$, and price varies inversely with output.

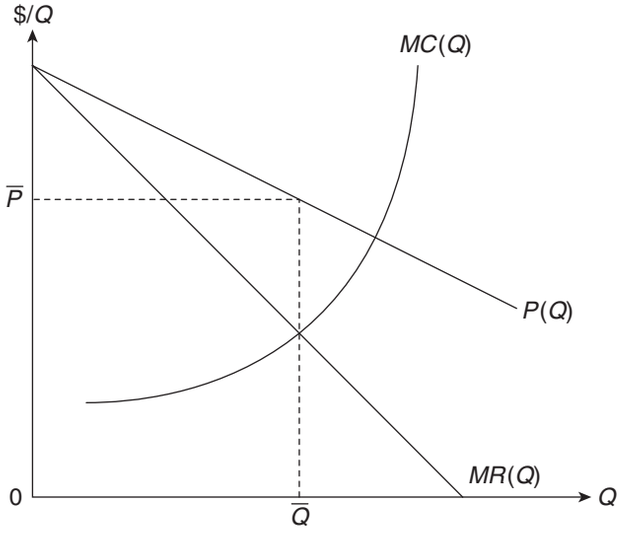


**Figure 11.8** Monopoly Profit-Maximizing Output and Price

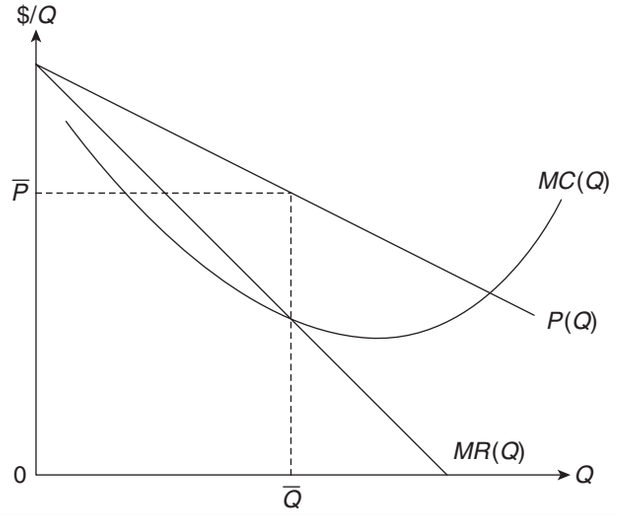


**Figure 11.9** Monopoly Profit Maximization With Falling Marginal Cost

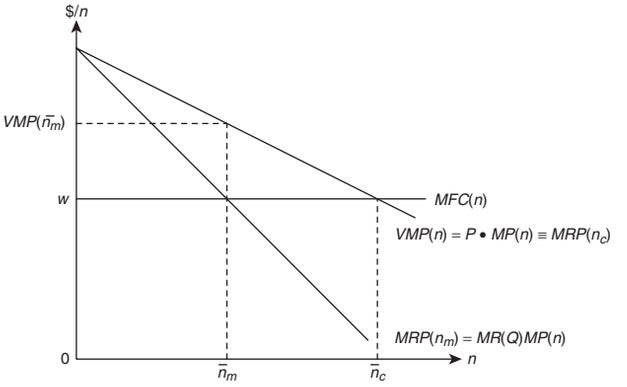


**Figure 11.10** Competitive Firm and Monopoly Profit-Maximizing Labor Input

The monopolist's marginal revenue is less than price. If the monopolist increases output, for example, this will lower the price. The change in revenue is the sum of the additional revenue from the additional output at the lower price and the lost revenue on the original output that it sold at the higher price. The monopolist's marginal revenue is

$$MR(Q) = \frac{\Delta TR(Q)}{\Delta Q} = P(Q) + \frac{\Delta P(Q)}{\Delta Q}Q < P(Q), \qquad (52)$$

because $\Delta P/\Delta Q < 0$. For example, with a linear inverse demand curve $P(Q)$ in Figure 11.8, $MR(Q)$ falls faster than $P(Q)$ as $Q$ increases and $P(Q) > MR(Q)$.

Unlike the price taker, a monopoly may maximize profit with $MC(q)$ falling. For example, in Figure 11.9, the firm's profit-maximizing output is $\overline{Q}$, where $MR(\overline{Q}) = MC(\overline{Q})$, $\Delta MC(Q)/\Delta Q > \Delta MR(Q)/\Delta Q$ (both slopes are negative), and $MC(q)$ cuts $MR(q)$ "from below."

## Monopoly Profit Maximization in the Labor Market

Monopoly in the product market affects the labor market as well as the product market. A monopolist in the product market need not have market power in the labor market. For example, the only barbershop in a small town has no effect on the market wage for barbers. For simplicity, we will assume a monopoly in the sale of its output that is a price taker in the labor market.

### *Marginal Revenue Product and Value of the Marginal Product*

We saw above that for a price taker in the product market, $P \equiv MR(q)$, and $MRP(n) \equiv VMP(n)$ in the labor market. The price-taking firm thus maximizes its profit by hiring labor to the point where $w = MRP(\overline{n}) = VMP(\overline{n})$.

For the monopoly, however, $MR(Q) < P(Q)$, which means that the monopolist's *MRP* is less than *VMP:*

$$MRP(n_m) = MR(Q)MP(n_m) < P(Q)MP(n_m) = VMP(n_m). \qquad (53)$$

In Figure 11.10, for example, note that the monopoly's $MRP(n_m)$ falls faster than $VMP(n_m)$ as it employs more labor.

### *Choice of Labor Input*

If the monopoly in the product market is a price taker in the labor market—that is, $MFC(n) \equiv w$—then $MRP(n) \equiv w$ and $VMP(n)$ are both linear and are both less than $ARP(n)$. If this were a perfectly competitive firm in the product market, it would hire $\overline{n}_c$, where $w = MRP(\overline{n}_c) \equiv VMP(\overline{n}_c)$. However, with monopoly in the product market, it will hire only $\overline{n}_m$,

where $w = MRP(\overline{n}_m) < VMP(\overline{n}_m)$. In other words, other things equal, a profit-maximizing monopoly will hire less labor to maximize profit than will a price taker in the product market.

## Monopoly Profit-Maximizing and Deadweight Loss

*Deadweight Loss in the Product Market*

The price-taking firm in a perfectly competitive product market maximizes profit by producing where the marginal benefit of the last unit of output to consumers just equals the marginal cost of producing it, as well as by employing labor where the value of the marginal unit of labor to the firm and to the consumers of the additional output is equal.

However, when the monopolist maximizes profit,

$$P(\overline{Q}) > MR(\overline{Q}) = MC(\overline{Q}), \tag{54}$$

which generates a deadweight efficiency loss from underproduction because

$$P(\overline{Q}) = MB(\overline{Q}) > MR(\overline{Q}) = MC(\overline{Q}). \tag{55}$$

In Figure 11.11, the efficient output and price are $\hat{Q}$ and $P(\hat{Q}) = MB(\hat{Q}) = MC(\hat{Q})$. This maximizes the combined net benefits to producers and consumers, or the sum of producer surplus and consumer surplus.

However, the monopoly maximizes profit by producing $\overline{Q}$ where $MR(\overline{Q}) = MC(\overline{Q})$ and charging a price of $\overline{P} = P(\overline{Q}) > MC(\overline{Q})$. At this price and output, $MB(\overline{Q}) > MC(\overline{Q})$ and there is a deadweight efficiency loss from underproduction equal to the shaded area in Figure 11.11. Unlike the price taker who must produce where $P = MC(\overline{q})$, a monopolist cannot maximize profit and produce the output that maximizes net benefits to producers and consumers,

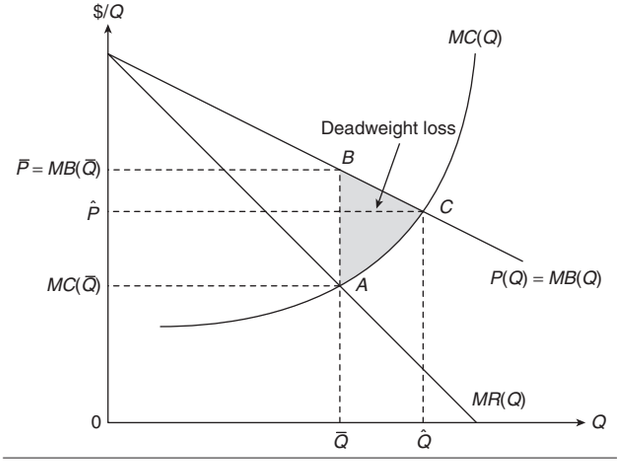


**Figure 11.11** Monopoly Equilibrium and Deadweight Loss

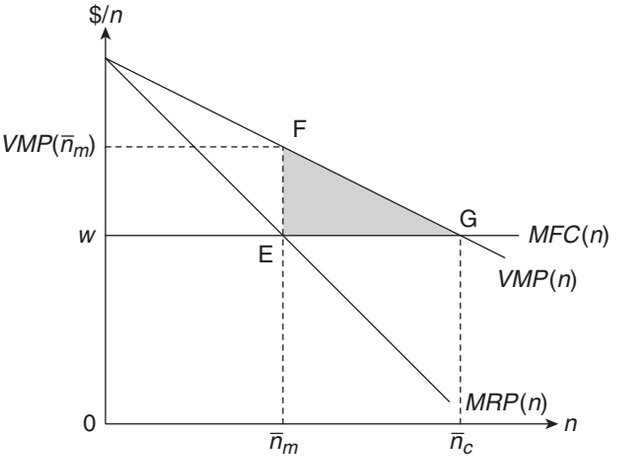


**Figure 11.12** Monopoly Labor Input and Deadweight Loss

and there is a conflict between the monopoly's interest in maximizing profit and society's interest in efficiency.

*Deadweight Loss in the Labor Market*

We know that price takers in the product and labor markets will maximize profit by employing labor where $w = MRP(\overline{n}_c) = VMP(\overline{n}_c)$ in Figure 11.12. This is efficient because the value of the last unit of labor hired by the firm equals the marginal benefit to consumers of the additional output produced by the marginal unit of labor.

However, a monopolist in the product who is a price taker in the labor market maximizes profit by employing labor to the point where

$$w = MRP(\overline{n}_m) < VMP(\overline{n}_m), \tag{56}$$

as shown in Figure 11.12. Because $VMP(\overline{n}_m) > MFC(\overline{n}_m)$, the additional benefits to consumers exceed the additional costs of employing more labor, and there is a deadweight loss of net benefits from underemployment at the monopoly's profit-maximizing employment of labor, equal to the shaded area *EFG* in Figure 11.12.

## Market Power and Profit Maximizing: Labor Market Monopsony

In labor markets in which an individual firm employs a sufficiently large share of the total employment that its employment choices affect the wage, the individual firm is not a price taker in the labor market. The simplest such case to analyze is a monopsony, in which there is only one firm in the labor market. The classic historical examples were the infamous "company towns" in isolated mining and lumber areas, in which a single firm employed all of the labor in the local labor market. Most of the "company

towns" are gone now, but there are still some markets in which individual firms or cartels of firms have considerable power over the wage that they pay—for example, professional sports teams and leagues, the NCAA in scholarship collegiate sports, universities in the hiring of college faculty, and "contractors" of undocumented alien workers.

*Wage and Marginal Factor Cost*

Like the price taker in the labor market, the necessary condition for maximizing profit by a monopsony firm in the labor market is $MFC(\overline{N}) = MRP(\overline{N})$, where $N$ is the total labor employed by the monopsony, the only employer in the labor market. However, the monopsony faces the market supply of labor and does not "take" a constant wage.

As in the monopoly model, it is easier if we consider the labor supply curve as $w(N_s)$ instead of $N_s(w)$. Assuming a positively sloped labor supply curve, by far the most plausible case, $\Delta w(N_s)/\Delta N > 0$. This means that the monopsony's $MFC(N)$ is

$$MFC(N) = w + \frac{\Delta w}{\Delta N} N > w(N), \tag{57}$$

which is shown graphically in Figure 11.13.

*Profit-Maximizing Input*

As a profit maximizer, the monopsonist will employ $\overline{N}_M$, where $MFC(\overline{N}_M) = MRP(\overline{N}_M)$. However, it will pay the supply price of this labor input, $w(\overline{N}_M)$—that is,

$$w(\overline{N}_M) < MFC(\overline{N}_M) = MRP(\overline{N}_M). \tag{58}$$

This result is shown graphically in Figure 11.13.

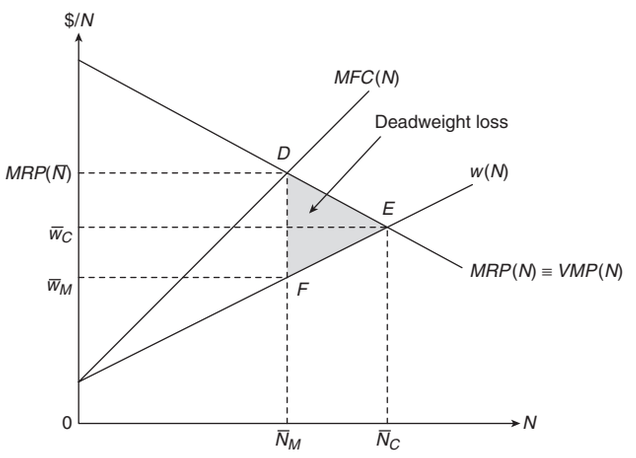


**Figure 11.13** Monopsony Profit-Maximizing Labor Input, Wage, and Deadweight Loss

If this were a competitive labor market, individual firms would maximize profit by hiring where $w \equiv MFC(n_c) = MRP(\overline{n}_c)$ on their demand curves for labor, and aggregate employment would be $\overline{N}_c$ at the competitive wage $w_c = MRP(\overline{N}_c)$. Thus, we would predict that a monopsony firm will hire less labor and pay a lower wage than the competitive equilibrium employment and wage.

*Monopsony and Deadweight Loss*

Labor in a monopsony market is paid a wage below its marginal revenue product, or the firm pays a wage less than the contribution of the marginal unit of labor to the firm's revenue. The difference between *MRP* and the wage is sometimes called monopsonistic "exploitation," which has strong normative implications of economic justice, but there are important efficiency implications as well.

If we assume for now that the monopsonist in the labor market is a price taker in the product market, then $MRP(N) \equiv VMP(N)$. The competitive employment and wage are efficient because the values of the marginal unit of labor to firms and consumers of output are equal. However, the monopsonist pays $w(\overline{N}_M) < MRP(\overline{N}_M)$, and there is a deadweight efficiency loss of area *DEF* in Figure 11.13.

*Bilateral Monopoly Profit Maximization*

A firm may have market power in product and factor markets. The simplest case is a monopoly in the product market that is also a monopsony in the factor market, or a bilateral monopoly. Like other profit-maximizing firms, the firm will produce where $MR(\overline{Q}) = MC(\overline{Q})$ and employ labor where $MRP(\overline{N}) = MFC(\overline{N})$.

However, because it is a monopolist in the product market, $MR(Q) < P(Q)$, therefore $MRP(N) < VMP(N)$ for a bilateral monopoly. As a monopsonist in the factor (labor) market, its $w(N) < MFC(N)$. The bilateral monopolist maximizes profit, then, by employing labor and paying a wage at which

$$w(\overline{N}) < MFC(\overline{N}) = MRP(\overline{N}) < VMP(\overline{N}). \tag{59}$$

This pushes the profit-maximizing employment and wage even lower than with monopoly or monopsony alone, widens the gap between $VMP(\overline{N})$ and $w(\overline{N})$, and increases the deadweight efficiency loss.

Figure 11.14 compares the results of profit maximization in the labor market in a competitive economy, monopsony in the factor market, and bilateral monopoly with monopsony in the factor market and monopoly in the product market. In the bilateral monopoly, the firm maximizes its profit by employing $\overline{N}_M'$ units of labor and paying a

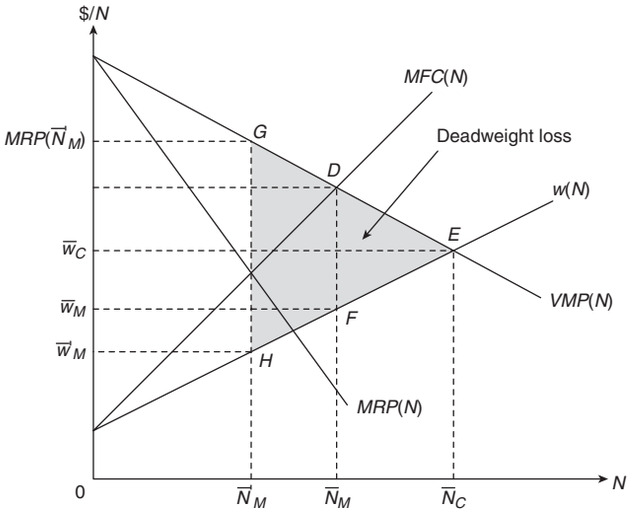


**Figure 11.14** Bilateral Monopoly, Labor Input, Wage, and Deadweight Loss

wage of $w\left(\overline{N}_M'\right)$. This is a lower wage and level of employment than the competitive case where $\overline{N}_c$ units of labor are employed at a wage of $w_c$. Employment and the wage are also lower in the bilateral monopoly than in the monopsony market, in which $\overline{N}_M$ units of labor are employed at a wage of $w(\overline{N}_M)$.

The deadweight efficiency loss in a bilateral monopoly combines the deadweight loss from monopoly in the product and from monopsony in the labor market. In Figure 11.14, for example, the bilateral monopoly deadweight loss is the area *GHE*. With monopsony alone, the deadweight loss is the area *DEF* from Figure 11.14.

## Why Profit Maximization?

Profit maximization is a plausible assumption to make about the motives of the decision makers who make the firm's choices. It is reasonable to assume that private firms are interested in their profits and that higher profits are better than lower profits. However, there is a longstanding controversy and large economic literature criticizing the profit-maximizing assumption as a highly unrealistic and simplistic picture of the operation and objectives of firms in modern industrial and commercial economies.

The people who make decisions in firms generally have multiple objectives in addition to profit. Proprietors of small businesses are motivated by their utility, which includes profit but also includes consuming leisure. The stockholders who own corporations are interested in the growth of the firm, the firm's share of its market, and the value of their shares in addition to the firm's profit.

Although the holders of voting stock own corporations, corporations are controlled by boards of directors and professional managers. Corporate managers may be more interested in maximizing their own utility, subject to the interests of the stockholders. If the performance of a corporation is unsatisfactory to stockholders, the stockholders may sell their shares, and depressed share prices will make it difficult for the firm to secure the necessary credit to conduct business.

Management in large corporations is conducted through a complex managerial bureaucracy in which lower level managers are generally more interested in satisfying their superiors and ultimately the stockholders than the corporation's profit. Simple profit maximization does not capture the complexity of decision making in large corporations.

Even among economists who accept the broad assumption that firms are maximizers, there are a number of alternatives to the assumption that they seek only to maximize profit. For example, Professor William Baumol (1967) constructed an analytical model in which the firm seeks to maximize its sales revenue, subject to a minimum profit constraint.

Some economists question whether firms, particularly corporations, actually *maximize* anything. Behavioral theories of the firm generally assume that managers are not maximizers but *satisficers.*

Modern economists in the "Austrian" tradition generally argue that while firms may seek maximum profit, they seldom if ever reach the goal because incomplete and diffused information would be required for maximizing profit. Rather, they argue that entrepreneurs are constantly "groping" for unexploited profit opportunities. They also argue that the simple assumption of profit maximization, particularly in perfectly competitive markets, understates or eliminates the active role of entrepreneurship in the choices that firms make.

Why, then, does mainstream microeconomics stick with the profit maximization assumption at all? The main reason is that most of the alternative models of the firm based on more extensive and realistic assumptions about the firm are much more complex than the profit-maximizing model, and most do not generate simple empirically testable explanations and predictions. Milton Friedman (1953) argued convincingly that as long as the firms behave "as though" they are profit maximizers, the profit-maximizing assumption is useful, even though it tells us little about the actual processes by which an actual firm makes its decisions.

Also, some of the "other" variables that motivate the firm's choices are directly related to profit. Share prices and the value of the firm tend to vary directly with profit. When firms fail to make stock analysts' predicted profit, or when the firms themselves project lower earnings and profits, this tends to depress share prices. Similarly, high earnings and profits tend to raise share prices and the value of the firm. Managers who are interested in protecting high salaries and generous perquisites are more likely to achieve higher utility

in a profitable firm than in one that is failing or facing the prospects of failure or a buyout by another firm. If proprietors seek to maximize their utility subject to income, and their firms' profits are the proprietors' incomes, it follows that profit has a positive utility to the proprietors.

Profit maximization remains a key assumption in the economics of the firm, even with its unrealistic and simplified picture of decision making by those who control firms because of its simplicity and the fact that the general explanations and predictions hold up well empirically. However, it is not a complete or very realistic picture of how those who control the firm make choices in product and factor markets.

## References and Further Readings


Ashley, C. A. (1961). Maximization of profit. *Canadian Journal of Economics and Political Science/Revue canadienne d'Economique et de Science politique, 27,* 91–97.

Baumol, W. J. (1967). *Business behavior, value and growth.* New York: Harcourt, Brace & World.

Blois, K. J. (1972). A note on X-efficiency and profit maximization. *Quarterly Journal of Economics, 86,* 310–312.

Boland, L. A. (1981). On the futility of criticizing the neoclassical maximization hypothesis. *American Economic Review, 71,* 1031–1036.

Cohen, K. J., & Cyert, R. M. (1975). *Theory of the firm: Resource allocation in a market economy.* Englewood Cliffs, NJ: Prentice Hall.

Cyert, R. M., & Pottinger, G. (1979). Towards a better microeconomic theory. *Philosophy of Science, 46,* 204–222.

Friedman, M. (1953). *Essays in positive economics.* Chicago: University of Chicago Press.

Herendeen, J. B. (1975). *The economics of the corporate economy.* Port Washington, NY: Dunellen.

Hovenkamp, H. J. (2008). *Neoclassicism and the separation of ownership and control* (University of Iowa Legal Studies Research Paper No. 08–52). Retrieved from http://ssrn.com/abstract=1315003

Koplin, H. T. (1963). The profit maximization assumption. *Oxford Economic Papers, New Series, 15,* 130–139.

Marshall, A. (1961). *Principles of economics* (9th [Variorum] ed.). New York: Macmillan.

Ng, Y.-K. (1974). Utility and profit maximization by an owner-manager: Towards a general analysis. *Journal of Industrial Economics, 23,* 97–108.

Sen, A. (1997). Maximization and the act of choice. *Econometrica, 65,* 745–779.

# 12

# Imperfectly Competitive Product Markets

Elizabeth J. Jensen
*Hamilton College*

Any introductory course in microeconomics spends a considerable amount of time examining perfectly competitive markets. It is important to understand this model; it serves as a benchmark for examining other industry structures and the welfare consequences of moving away from perfect competition. However, it is also important to look at imperfectly competitive output markets—markets in which products are not perfectly homogeneous or in which there are only a few sellers. While the perfectly competitive model assumes a large number of buyers and sellers, each of which is a price taker, the monopoly model assumes the opposite: one seller with complete control over price. Structurally, most markets are neither perfectly competitive nor monopolistic; they fall somewhere in between these two extremes of the competitive spectrum. The "in-between" markets are classified as either monopolistic competition or oligopolies depending on the number of firms in the market and the height of barriers to entry and exit. We turn first to monopoly.

## Monopoly

A monopoly is the only producer of a good for which there are no close substitutes. This puts the monopolist in a unique position: It can raise its price without losing consumers to competitors charging a lower price. Thus, the monopolist *is* the industry and faces the downward-sloping market demand curve for its product. The monopolist can choose any point along that demand curve; it can set a high price and sell a relatively small quantity of output, or it can lower price and sell more output.

Very few—if any—industries in the real world are pure monopolies. Examples of industries that come close include public utilities such as the local distributor of electricity or natural gas, the cable company in most communities, and, in a small isolated town, the local grocery store or gas station. Nonetheless, this model highlights how a firm with market power chooses its output level and price and helps us understand the welfare consequences of this choice.

### Marginal Revenue

All firms, ranging from perfectly competitive firms to monopolies, maximize profits by producing the quantity of output for which marginal revenue equals marginal cost. Recall the definition of marginal revenue: the additional revenue a firm receives from selling one additional unit of output. Because a monopoly faces the market demand curve, the only way it can sell an additional unit of output is by lowering the price it charges on *all* units. Consequently, marginal revenue for a monopoly has two components: an output effect and a price effect. The monopoly gains revenue from the additional unit of output sold but loses revenue on all of the units previously sold at a higher price. Marginal revenue incorporates both the gain and the loss; Figure 12.1 shows this trade-off.

At price $P_1$, the monopoly sells $Q_1$ units of output, and total revenue is the area of the rectangle $OP_1AQ_1$. To sell an additional unit of output, the monopoly has to reduce price to $P_2$, leading to total revenue of $OP_2B(Q_1 + 1)$. The monopoly gains rectangle $Q_1CB(Q_1 + 1)$ in total revenue; this area is equal to $P_2$. However, because of the lower price, the monopoly loses area $P_1AC\,P_2$, which is the initial quantity

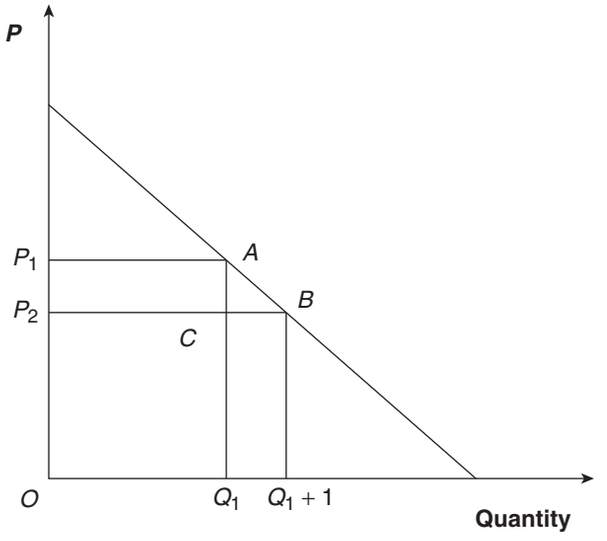


**Figure 12.1** Changes in a Monopolist's Total Revenue Resulting From a Price Cut

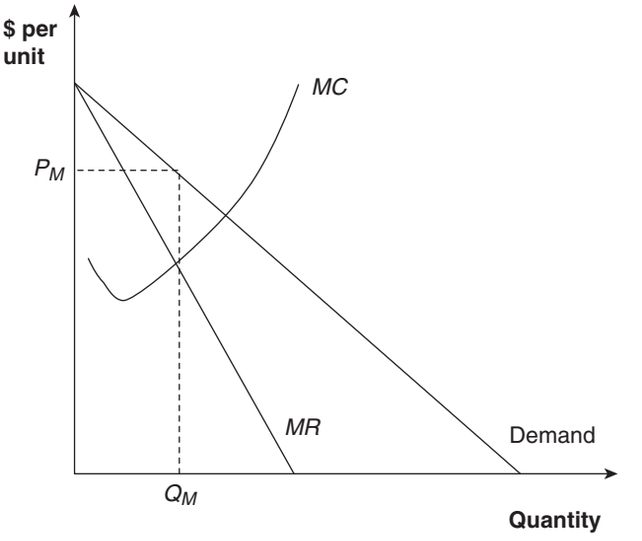


**Figure 12.2** Profit-Maximizing Quantity and Price for a Monopolist

sold times the change in price. Marginal revenue, the sum of the two, is thus always less than price for a monopoly. The precise relationship between marginal revenue and price is given by

$$MR = P + Q\frac{\Delta P}{\Delta Q}.$$

This can also be written as

$$MR = P\left(1 - \frac{1}{|\varepsilon|}\right),$$

where ε is the price elasticity of demand. Rewriting the relationship in terms of the elasticity of demand highlights the fact that even a monopolist must consider consumers' ability and willingness to change quantity in response to a change in price.

Graphically, the marginal revenue curve lies below the market demand curve at each level of output. While the exact relationship between marginal revenue and demand depends on the shape of the demand curve, there is a useful relationship between any linear demand curve and its marginal revenue curve. Specifically, the marginal revenue curve will also be linear, with the same vertical intercept as the demand curve but twice the slope. Consequently, the horizontal intercept of the marginal revenue curve is half the horizontal intercept of the demand curve. For example, if the market demand curve is given by $P = 160 - Q$, then the corresponding marginal revenue curve is given by $MR = 160 - 2Q$. This mathematical relationship, sometimes called the twice-as-steep rule, can be helpful in graphing demand and marginal revenue and in doing algebraic problems.

## Profit Maximization

Figure 12.2 shows that $Q_M$, where marginal revenue equals marginal cost, is the profit-maximizing quantity of output for the monopolist.

At any output level below $Q_M$, marginal revenue is greater than marginal cost, so producing an additional unit of output will increase profits; at any output level greater than $Q_M$, marginal cost is greater than marginal revenue, so the monopoly can increase profits by cutting back on quantity. Once the firm has chosen $Q_M$, it uses the market demand curve to determine the profit-maximizing price, $P_M$. Setting this price will enable the monopoly to sell its desired quantity of output. As the figure shows, price is greater than marginal cost for the profit-maximizing monopolist.

Continuing with the algebraic example, suppose the monopolist faces constant marginal and average costs of \$10. Setting marginal revenue, $160 - 2Q$, equal to marginal cost, we find that a profit-maximizing monopolist wants to sell 75 units of output. Substituting $Q = 75$ into the market demand curve, $P = 160 - Q$, shows that the monopolist will set price equal to \$85, considerably above the marginal cost of \$10.

## The Welfare Effects of Monopolization

Because the monopolist *could* have chosen to produce the perfectly competitive output level but did not, we would expect that monopolization of an industry makes the firm better off. However, because prices are higher and output is lower under monopoly than under perfect competition, we would also expect that consumers are worse off under monopoly. To see these effects and to look at total welfare, economists use consumer and producer surplus.

Figure 12.3 compares a market under perfect competition and under monopoly, assuming that costs are the same under both industry structures.

The perfectly competitive output is $Q_{PC}$; the perfectly competitive price is $P_{PC}$. Consequently, consumer surplus under perfect competition is area *ABC*, and producer surplus is area *CBD*. If the industry is monopolized, price

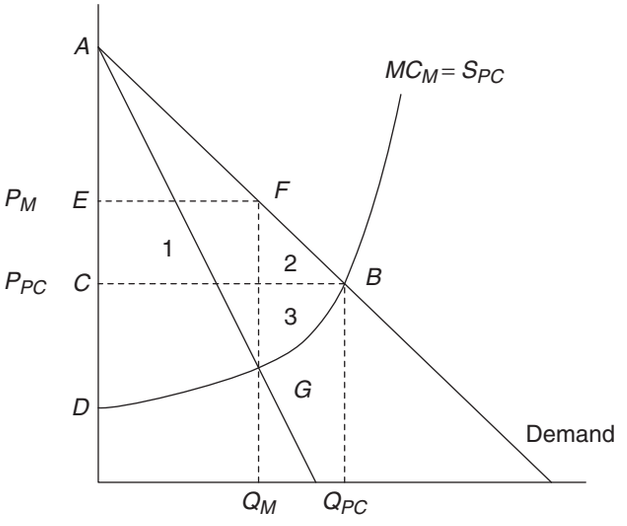


**Figure 12.3** Welfare Loss Due to a Monopoly

increases to $P_M$ and quantity decreases to $Q_M$. Consumer surplus is now area *AEF*, so consumers lose Areas 1 and 2. Producer surplus under monopoly is area *EFGD*, so producers gain Area 1 and lose Area 3. Because Area 1 is transferred from consumers of the good to the owner(s) of the monopoly, it is *not* a loss of welfare from society's point of view. However, Areas 2 and 3 are clearly losses to society, reflecting surpluses that would have gone to some group under perfect competition. This loss, called, the deadweight loss, measures the misallocation of resources due to monopoly.

The deadweight loss arises because the monopoly price is greater than the marginal cost of the good. Even though some consumer is willing to pay more for an additional unit of the good than it would cost to produce that unit, the monopolist does not produce it. This is the fundamental problem associated with monopoly power; understanding this helps us understand concern about any firm with market power and public policy toward such firms.

## Monopolistic Competition

In many industries, consumers do not view products of different firms as perfectly homogeneous—that is, products are differentiated. Sellers differentiate their products in many ways, including physical characteristics, location of stores, service, and warranties. Sellers also try to create subjective image differences, persuading consumers that their brand of shampoo, toothpaste, or laundry detergent is different from (and better than) other brands. A consumer chooses Herbal Essences shampoo instead of Dove or Pantene partly based on color, scent, packaging, and image (correct or incorrect) about how the shampoo will work. Because of product differentiation, some consumers will pay more for Herbal Essences shampoo.

Procter & Gamble, through its Clairol division, is the sole producer of Herbal Essences, giving it some monopoly power. This implies that the demand curve for Herbal Essences is downward sloping, in contrast to the perfectly horizontal demand curve faced by a seller in a perfectly competitive market. But the market power of a monopolistically competitive firm is limited because there are many producers of similar products. While a consumer might pay $0.50 or $1.00 more for a bottle of Herbal Essences, she is unlikely to pay $5.00 more. Thus, the downward-sloping demand curve for Herbal Essences is relatively elastic.

The second important characteristic of a monopolistically competitive market is easy entry and exit. If firms are making positive economic profits, new firms will develop their own brands and enter the market. Similarly, existing firms will leave the market if their products are unprofitable. Consequently, economic profits must equal zero when a monopolistically competitive industry is in long-run equilibrium.

### Equilibrium in the Short Run and in the Long Run

Figure 12.4(a) shows short-run equilibrium for a monopolistically competitive firm.

The profit-maximizing firm will produce $q_{SR}$, where marginal revenue equals marginal cost. In this case, the corresponding price, $P_{SR}$, is above average cost so the firm earns a positive economic profit in the short run.

As in perfect competition, positive economic profits cannot persist in the long run. The profits will encourage other firms to spend the funds necessary to develop, produce, and advertise new products. Over time, this entry will shift each existing firm's demand curve to the left until economic profits reach zero. Figure 12.4(b) shows the long-run equilibrium: The firm produces $q_{LR}$, where marginal revenue equals marginal cost and where price equals average cost. Because entry and exit are easy, no firm can make economic profits in the long run even though each firm has some market power.

### Welfare Effects of Monopolistic Competition

Economists have identified two sources of inefficiency in a monopolistically competitive market. First, note from Figure 12.4(b) that the firm produces a smaller output level than that which minimizes long-run average cost. This excess capacity indicates that if there were fewer firms in the industry, each could operate at a larger scale and a lower average cost. The second source of inefficiency results from the fact that in both the short run and the long run, price is greater than marginal cost. Consequently, as in monopoly, there is a deadweight loss in a monopolistically competitive market; each firm chooses to produce at an inefficiently small output level considering consumers' preferences.

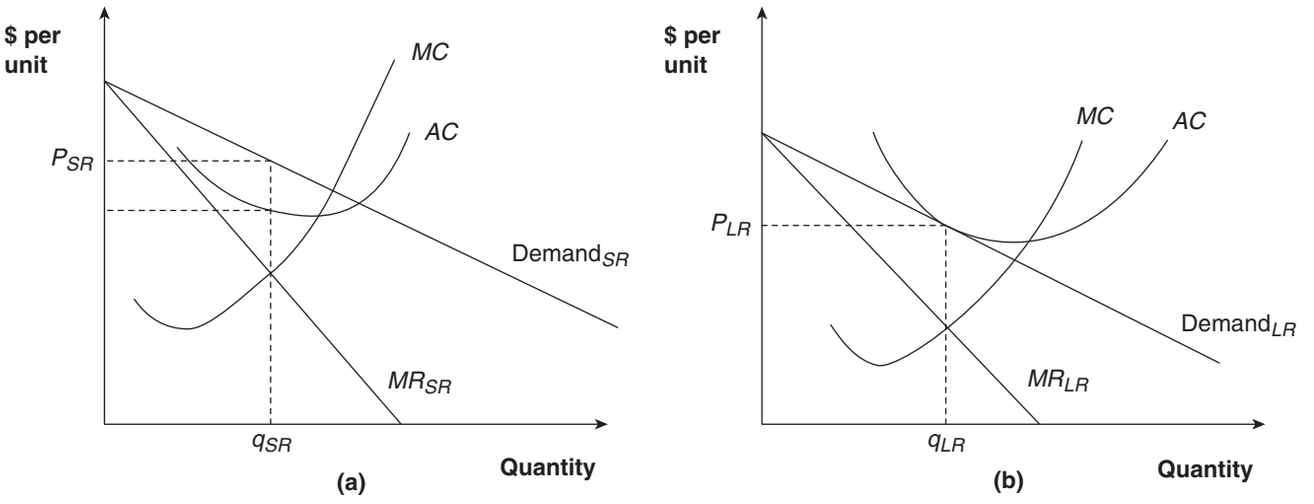


**Figure 12.4** (a) Short-Run and (b) Long-Run Equilibrium Under Monopolistic Competition

While these inefficiencies are costs to society, consumers undoubtedly value the opportunity to choose among a variety of products with different characteristics. Although it is hard to quantify the gains from product diversity, it is possible that they outweigh the welfare costs. This is especially likely when brands are reasonably substitutable; if no firm has much monopoly power, deadweight loss is small, and the firm is operating close to the minimum average cost.

## Oligopoly

An oligopolistic market consists of only a few producers. Furthermore, entry by new firms into the market is impeded; consequently, firms in an oligopolistic industry can earn substantial economic profits over the long run. Products produced by oligopolies can be relatively homogeneous, like steel, or differentiated, like automobiles. Many industries are oligopolistic, including aluminum, cigarettes, breakfast cereals, petrochemicals, computers, and tennis balls.

Predicting oligopolistic price and output or decisions about nonprice strategies such as advertising or investment decisions is complicated. Because of the small number of firms in the industry, oligopolists recognize their interdependence. Kellogg's understands that its actions affect General Mills and other manufacturers of breakfast cereals. Similarly, General Mills knows that what it decides matters for Kellogg's. Each firm realizes that any significant move on its part, such as a decision to cut price by 10% or to run a new advertising campaign, is likely to result in countering moves by its competitors. Firms therefore have to try to determine their competitors' probable responses as they make their decisions. Not surprisingly, these strategic considerations can be complex, suggesting that there is not just one model of equilibrium in an oligopoly, as there is in perfect competition, monopoly, or monopolistic competition. Factors such as whether firms are competing in prices or in quantities, the extent of information they have about their rivals, and how often they compete with each other matter in modeling choices in an oligopoly. Consequently, when examining the competitiveness of real-world oligopoly markets, it is important to combine theory with empirical evidence, remembering that "one size does not fit all." Before turning to specific models, we consider some of the important factors to think about.

### Competing in Quantities or Competing in Prices?

In some industries, it makes sense to think of firms choosing the price of their output and then selling as much as they can at that price. If they can easily produce an additional unit of output, firms will compete with each other based on price. Examples include the computer software, video games, and insurance industries. However, in other industries, firms have to choose their production capacity in advance, and quantity cannot be changed quickly or easily. Firms in these industries choose quantity with price adjusting to clear the market. Such industries include steel, automobiles, and passenger aircraft. As we will see, choosing prices instead of choosing quantities leads to a different oligopoly equilibrium.

### Collusion or Competition?

Another consideration is how oligopolists interact in their attempts to maximize profits. Perhaps they collude with each other, working to maximize joint profits and dividing up the market. If firms can see the "big picture," they will realize that collusion is more profitable in the long run than competing with and undercutting each other. However, any collusive agreement faces two types of problems. First, each individual firm is tempted to cheat on the

agreement to increase its short-run profits. And second, firms have to pay attention to the relevant law. In the United States, it is illegal to write a contract to keep prices high. Consequently, collusive, joint-profit-maximizing agreements are more stable in some industries than in others, depending on how easy it is to reach an agreement without overt collusion and how quickly a firm can increase output if it tries to take sales from a rival.

## Simultaneous or Sequential Decision?

Sometimes oligopoly firms must choose their strategies simultaneously or at least close enough together in time that they are deciding without knowing their competitors' decisions. For example, Sears and Best Buy choose prices for flat-screen TVs without knowing what price the other store has chosen that day. Coca-Cola and Pepsi decide on new advertising campaigns without seeing each other's plans. However, in some situations, one firm acts first, and then other firms decide how to respond. Sequential business decisions include entry decisions in which potential competitors must decide whether to enter a particular market and bargaining decisions such as those in which a firm negotiates with its suppliers over prices. In modeling oligopoly behavior, it is important to think about whether the particular interaction is better thought of as simultaneous or sequential.

## Equilibrium in an Oligopolistic Market

To think about market price and output in an oligopoly, we must first consider what an equilibrium would look like when firms are explicitly taking each other's behavior into consideration. The principle used in many models of oligopoly behavior is the Nash equilibrium, named in honor of the mathematician John Nash who developed it. When an industry has reached a Nash equilibrium, each firm is doing the best that it can given the choices made by its competitors. In this equilibrium, no firm has an incentive to change its strategic choices unilaterally. This concept is a powerful one; with it in hand, we can turn to some common models of oligopoly. To keep things simple, we will focus on duopoly markets, those in which two firms are competing with each other; these results do generalize to markets with more than two firms, although the math becomes more complicated.

## Models of Oligopoly

### *The Cournot Model*

First introduced by the French economist Augustin Cournot in 1838, the Cournot model has turned out to be a powerful model, applicable in many situations. In its simplest form, it depicts a duopoly; the firms produce a homogeneous product, have the same marginal costs, and know the market demand curve. Firms in this model are choosing quantity and are choosing simultaneously and independently.

To be concrete, let us use the market demand from the discussion of monopoly, $P = 160 - Q$; assume that each duopolist operates with constant marginal and average cost of 10, as the monopolist did. Each firm maximizes profits, taking into account the amount of output it expects its rival to produce. Thus, Firm 1 sees its demand curve, called the residual demand curve, as $P = 160 - q_2 - q_1$. Because Firm 1 thinks of the quantity produced by Firm 2 as constant, its marginal revenue curve is given by $MR_1 = 160 - q_2 - 2q_1$, derived using the twice-as-steep rule from monopoly. As always, profit maximization requires setting marginal revenue equal to marginal cost. Doing this and solving for $q_1$ gives us

$$q_1 = \frac{(150 - q_2)}{2}.$$

This relationship, called the reaction function, shows the profit-maximizing output level for Firm 1 given any output level chosen by Firm 2. So, for example, if Firm 1 sees that Firm 2 is producing 10 units of output, it would choose to produce 70 units of output.

Similar reasoning leads to the reaction function for Firm 2, given by

$$q_2 = \frac{(150 - q_1)}{2}.$$

Because each firm chooses its output level according to its reaction function, the Cournot equilibrium is found at the intersection of the two. Solving the two equations together for the example gives equilibrium quantities of 50 units of output for each firm. Note that this equilibrium is a Nash equilibrium: Each firm is maximizing its profits given the level of output produced by its competitor, so neither firm wants to change its output level. With total output of 100 units, the market price in equilibrium is \$60. Profits, calculated using the formula $(P - AC) \times q$, are thus \$2,500 for each firm, so total industry profits are \$5,000.

To get some perspective on this equilibrium, it is useful to compare it to both perfect competition and monopoly. Under perfect competition, price would equal marginal cost, \$10; total output would be 150; and no firm would make economic profits. As we saw earlier, the monopoly output is 75 and the market price is \$85; monopoly profits are (\$85 − \$10) × 75, or \$5,625. Thus, we see that the Cournot duopolists produce less output and make considerably more profits than competitive firms; we also see that they do not do as well as colluding firms working together to produce the monopoly outcome.

### *The Stackelberg Model*

Suppose now that one of the duopolists can choose its output level first; its competitor then observes this output level and reacts to it. While the Cournot model makes sense for an industry composed of basically similar firms, the Stackelberg model might be more appropriate for an industry in which one firm typically is the leader in introducing new products or determining output levels. For

example, IBM has been the leader in the mainframe computer market for many years.

Let us continue the example, assuming that Firm 1 is the leader and Firm 2 is the follower in the industry. Because Firm 2 chooses after Firm 1, it treats Firm 1's output as given and will use the Cournot reaction function,

$$q_2 = \frac{(150 - q_1)}{2},$$

to select its output level. The leader, however, can take advantage of its leadership position to do better than it would do under Cournot. The leader uses its knowledge of the follower's reaction function in making its own decision:

$$P = 160 - q_2 - q_1 = 160 - \left(\frac{150 - q_1}{2}\right) - q_1$$
$$= 160 - 75 + \frac{q_1}{2} - q_1 = 85 - \frac{q_1}{2}.$$

Using the twice-as-steep rule, the corresponding marginal revenue for Firm 1 is given by MR = 85 – $q_1$; setting this equal to marginal cost of $10 tells us that the profit-maximizing quantity for the Stackelberg leader is 75. The follower will then produce 37.5 units of output for a total industry quantity of 112.5; the market price will be $47.5. As expected, Firm 1's profits, $2,812.50, are better than under the Cournot model. However, Firm 2's profits under Stackelberg fall to $1,406.25; indeed, the follower does only half as well as the leader in this model.

The result that the leader does considerably better than the follower is an important feature of the Stackelberg model. Remember that the two firms produce identical products and face the same, constant marginal cost, so the advantage for the leader is not a result of product differentiation or better technology. In addition, Firm 2 does worse in Stackelberg despite the fact that it has concrete information about Firm1's choice—information it does not have in the Cournot model. The key difference between the two models is that Firm1's Stackelberg choice to produce a large quantity of output (75 units in the example) is irreversible. To maximize profit, the follower has no choice other than to take this large output level as given and produce a smaller quantity of output itself. Stackelberg's modification to Cournot's initial model is important, showing us that moving first clearly has advantages.

*The Bertrand Model*

A different modification of Cournot's model, developed by another French economist, Joseph Bertrand, in 1883, looks at the equilibrium that results if oligopoly firms choose price rather than quantity. As in Cournot's model, firms produce a homogeneous product and make their decisions simultaneously. It turns out that choosing price instead of quantity has a dramatic effect on the outcome.

In our example, each firm has a constant marginal cost of $10; market demand is $P = 160 - Q$. Because products are identical, the firms realize that if their prices differ, consumers will buy from whichever firm has the lower price. If they charge the same price, consumers will be indifferent between the two, so the firms will split the market.

Imagine that you are deciding on price for one of the two firms. If you expect your competitor to set price anywhere above marginal cost, you have an incentive to set price just below the competitor's price and take the entire market. But of course, your competitor reasons similarly, hoping to undercut your price. Consequently, the only Nash equilibrium for the simple Bertrand model is the competitive equilibrium. Each firm will set price at marginal cost, $10, producing 75 units and making zero economic profit. At this point, neither firm wants to change its price: Raising price would lead to losing all sales to the rival, and lowering price below $10 would result in losses on every unit of output.

Economists have shown that introducing product differentiation, cost differences across firms, or capacity constraints to the Bertrand model results in an equilibrium price above marginal cost. Still, the Bertrand model shows the importance of the choice of the strategic variable. Choosing price instead of quantity leads to a strikingly different equilibrium, illustrating again the importance of looking carefully at the way any real-world oligopolistic industry works when thinking about its competitiveness.

## Game Theory

As we have seen, analyzing oligopoly behavior is complicated because of the strategic interactions among the firms. Economists have adopted tools in game theory, developed initially as ways to think about military strategy during World War II, to analyze these interactions. Indeed, although the models we have already considered were developed before the pioneering work in game theory, economists have shown that they fit within this theoretical framework.

Game theory has two branches. Cooperative game theory assumes that firms work together to maximize joint profits, dividing up the market among them. In noncooperative game theory, firms act independently, attempting to maximize individual profits while taking their rivals into account. We turn first to this type of game, looking at a classic game theory example that illustrates the problem faced by oligopolists.

*Simultaneous Games: The Prisoner's Dilemma*

Suppose that BASF and Merck are the only producers of vitamin C and that they are choosing quantity in a simultaneous game. Also, to keep things simple, assume that each firm can pick only one of two alternative quantities: 20 million tons or 30 million tons. Given these simplifying assumptions, the outcomes of combinations of choices can be represented in a payoff matrix containing four cells, shown as Figure 12.5.

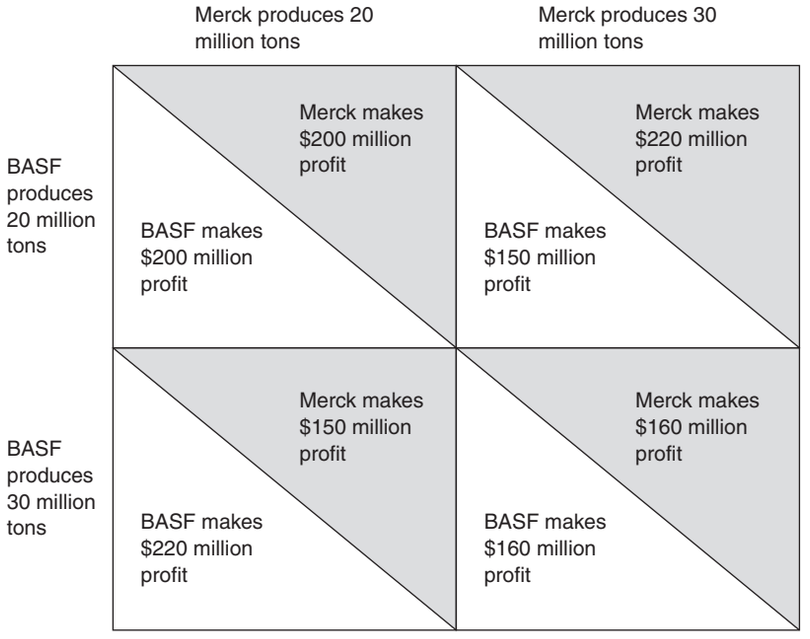


**Figure 12.5** Payoff Matrix for Firms Choosing Quantity of Output

The payoff matrix shows that joint profits of the two firms are maximized if each produces 20 million tons of vitamin C; this would be the cooperative outcome. However, imagine that you are deciding on production for BASF. If you expect Merck to produce 20 million tons, you can earn higher profits by cheating on the collusive agreement and increasing your output to 30 million tons ($220 million in profits is better than $200 million in profits). On the other hand, if you expect Merck to increase its output to 30 million tons, then you also want to produce the larger output level ($160 million in profits is better than $150 million). Of course, decision makers at Merck reason the same way; producing 30 million tons is called a dominant strategy because it is the best choice regardless of the other firm's actions. Thus, the Nash equilibrium is for both firms to produce 30 million tons of vitamin C even though each ends up with lower profits than at the cooperative outcome. This example helps us understand why it is often difficult for oligopolists to maintain an agreement about quantity, price, or other strategic variables.

This outcome is frustrating for the two firms. If they had some way of enforcing cooperative behavior, such as signing an unbreakable contract, they could avoid the prisoner's dilemma equilibrium. However, even if the firms could figure out how to write an enforceable agreement, they would be subject to antitrust law, which holds such agreements illegal. Thus, it seems as if the undesirable noncooperative equilibrium, with both firms producing 30 million tons of vitamin C, is the inevitable outcome. Might anything change this?

### *Overcoming the Prisoner's Dilemma?*

If firms are playing a prisoner's dilemma game only once, they will arrive at the noncooperative Nash equilibrium. But most oligopolists interact with each other repeatedly. Using our simple example, BASF and Merck regularly have to decide how much vitamin C to produce. Consequently, in addition to thinking about short-run profits, managers are likely to consider the long-run implications of a decision to "cheat" on their reputation and on other firms' reactions. Each firm may reason that production decisions in the current period are a signal about intentions in future periods, understanding that it is actually in its own long-run best interest to help the other by restricting output. This resolution of the prisoner's dilemma can thus occur even without explicit collusion or an enforceable agreement about strategy; economists call such behavior tacit collusion.

Tacit collusion is successful in some oligopolistic industries but not in others. Looking at real-world oligopolies, we can identify several key factors that complicate life for oligopolists, making it more difficult to collude tacitly or undermining the stability of any arrangements:

- *Large number of firms.* Suppose that the China Pharmaceutical Group joins BASF and Merck in the vitamin C industry. With three firms in the industry instead of two, the reward to any firm for cheating—the gain in short-run profits—increases compared to the gain with only two firms. Also, with more firms, it becomes harder to monitor rivals' actions. As the number of firms in an industry increases, tacit collusion is less likely to be successful.
- *Low industry concentration.* Consider two industries, each with 10 firms; in Industry 1, the two largest firms control 80% of the market, while the combined market share of the two largest in Industry 2 is 20%. In Industry 1, the two largest firms can behave as price leaders with the smaller firms following, making coordination easier.

On the other hand, it is more difficult to know who will lead on price changes, advertising campaigns, or other strategic decisions in Industry 2, where market share is more evenly divided.

- *Cost asymmetries.* If firms have different costs, it is difficult to agree on output and price. High-cost firms would prefer a higher price and lower combined output; low-cost firms would prefer greater industry output and a lower price. Also, differences in costs across firms affect firms' incentives to cheat on any agreement: Low-cost firms have more incentive to cheat than high-cost firms because they will gain more in profits by expanding output.
- *Product differentiation.* Vitamin C is a fairly homogeneous product, so producers would need to coordinate only on a single price or level of output. On the other hand, consider an industry such as the beer industry; coordination in this industry would require a schedule of prices for products ranging from India pale ale to stout to pumpkin ale. The more differentiated products are, the more possibilities there are for disagreement and difficulty in recognizing whether rivals are cheating on an agreement.
- *Rapid rate of innovation.* Not surprisingly, tacit collusion is more difficult in industries experiencing a high rate of technological change than in industries in which product characteristics are stable. Consider, for example, the cell phone industry and the rapid evolution of products within this industry. The frequent changes in product characteristics, production costs, and demand complicate life for oligopolists.

### *Sequential Games*

In sequential games, players move in turns. The Stackelberg model, in which one firm sets output before the other, is an example of a simple sequential game. Other examples include capacity investment by an incumbent firm and a potential competitor's decision about whether to enter a market or a new advertising campaign by one oligopolist and its rivals' responses. Firms may also make decisions about adopting new technology sequentially. As we saw in the Stackelberg model, changing the setup of a strategic game from simultaneous to sequential affects the outcome. Here we examine a simple entry-deterring game to illustrate solution concepts for sequential games.

Suppose that a small town in New Mexico currently has no drugstore. Executives at CVS have decided to enter this market and must decide what size store to build. They know that the opportunity cost of the funds necessary to build the store is 12%—that is, the store must earn a rate of return of at least 12% to be a worthwhile investment. If CVS builds a small store in the town, it will earn economic profits with a rate of return of 25%. Because of higher costs, the rate of return for a larger store will be only 18%.

Based on this information, it seems obvious that CVS should build the smaller store. However, the executives at CVS know that Kinney Drugs is also considering entering this market. If CVS builds a small store and Kinney enters, each will earn a 15% rate of return. If CVS builds a large store and Kinney enters, both will have to cut prices; each firm will earn only 10%, thereby incurring economic losses.

Sequential games are analyzed using a decision tree, such as that shown in Figure 12.6.

Each box or decision node represents a point where one of the firms must make the decision shown in the box. CVS makes the initial decision about size of store, and then Kinney responds by deciding whether to enter the

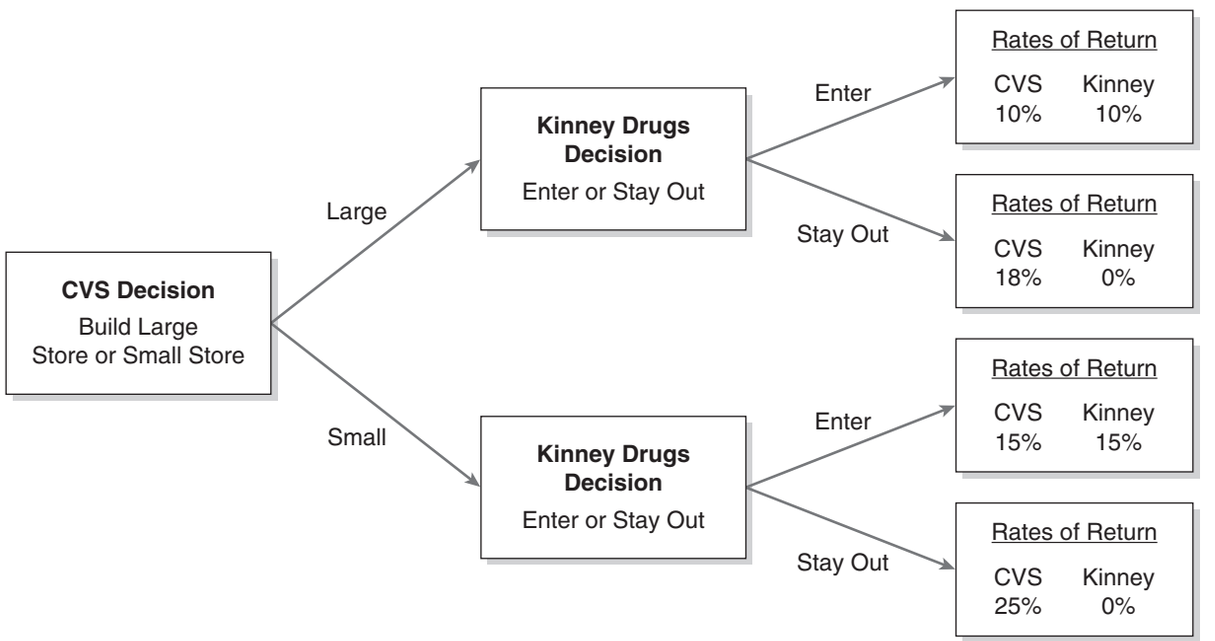


**Figure 12.6** Deterring Entry by Capacity Decision

market. The boxes on the right-hand side show the rates of return resulting from each combination of choices.

To solve a sequential game, firms should look ahead and then reason back. From the rates of return, CVS can reason that if it builds a small store, Kinney will enter, lured by the prospect of a 15% rate of return. On the other hand, if it builds a large store, Kinney will not enter the market: A 10% rate of return is below the opportunity cost of the funds. Thus, CVS understands that its two possible outcomes are rates of return of 18% or 15%; the profit-maximizing choice is to build the large store, deter entry, and earn 18%. Even though CVS would like to earn 25%, this outcome is not attainable as long as Kinney is behaving rationally.

Executives at CVS might well be frustrated by this and wonder if a *threat* to expand its store in the event of entry by Kinney Drugs would be enough to deter entry. Figure 12.7 shows the game tree for this scenario.

In this situation, CVS builds a small store but announces that it will expand its store if Kinney enters the market. Using the solution technique of looking ahead and reasoning back, we see that CVS's threat is not credible. Once Kinney Drugs has entered the market, CVS would be shooting itself in the foot by expanding, earning a rate of return of 10%—below the opportunity cost of its funds—instead of making positive economic profits at a 15% rate of return. The only equilibrium for this scenario is for Kinney to enter the market and for CVS to accommodate the entry by staying small.

This simple example illustrates the use of decision trees to help firms predict the actions of rivals and think through the implications of strategies. It also highlights the importance of *credibility*. An incumbent firm playing a sequential game must often make a *commitment*—to additional capacity, to a particular advertising strategy, to a new product—to change a strategic choice by potential or actual rivals. Economists have extended the analysis of sequential games to situations in which games are repeated, information is uncertain, and incumbents can be weak or strong, but the basic solution principle of looking ahead and reasoning back still applies. Decision trees are widely used in business planning.

## Limiting Market Power: The Antitrust Laws

Market power on the part of firms reduces consumer surplus and creates deadweight loss. Because of concerns about these adverse effects, governments in most countries have enacted policies related to market structure and behavior of firms. In the United States, Congress passed the first federal antitrust law, the Sherman Antitrust Act, in 1890. This act resulted from political pressures associated with the changing organization of American industry. Before the Civil War, high transportation costs kept the geographic size of most markets relatively small. After the Civil War, several factors led to the growth of national industrial markets: mass production techniques, a national railroad system, the development of modern capital markets, and the liberalization of the laws of incorporation in several states. As national markets arose, many small regional manufacturers faced increasing competition; farmers also were unhappy about what they saw as the abuses of firms with market power. Both the Democrats and the Republicans included planks in their party platforms opposing the trusts during the 1888 election, and Senator John Sherman introduced the first antitrust bill in the Senate shortly before the election. The bill was signed into law on July 2, 1890.

The Sherman Act has two substantive sections. Section 1 states in part, "Every contract, combination in the form of a trust or otherwise, or conspiracy, in restraint of trade of commerce among the several states, or with foreign

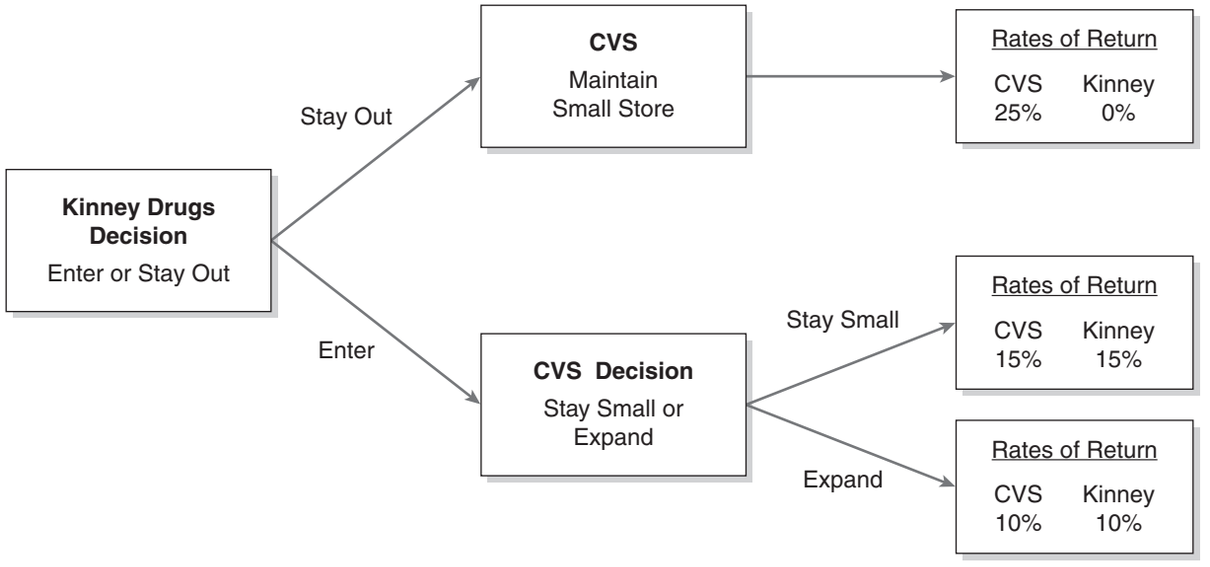


**Figure 12.7** Examination of Credibility of a Threat

nations, is hereby declared to be illegal." Section 2 includes language that "every person who shall monopolize, or attempt to monopolize, or combine or conspire with any other person or persons to monopolize any part of the trade or commerce among the several States, or with foreign nations, shall be deemed guilty of a misdemeanor."

Although this language sounds quite strong, the legal interpretation and implications led to problems soon after the law was passed. By making it illegal for rival firms to collude with each to fix prices, Section 1 paradoxically strengthened the incentive of competitors to merge. In part, this act led to the first major merger wave in the United States at the turn of the century. The wording of Section 2 of the Sherman Act led to some confusion about what, in fact, constituted a violation. In a famous ruling in 1911 (the *Standard Oil* decision), the Supreme Court annunciated a precedent that has come to be known as the Rule of Reason. According to this interpretation, only *unreasonable* attempts to monopolize are illegal; the mere possession of market power is legal as long as it is not a result of aggressive business tactics aimed at obtaining a monopoly.

In 1914, Congress responded to concerns about the Rule of Reason and about some specific business practices by passing two additional major pieces of antitrust legislation, the Clayton Act and the Federal Trade Commission Act. The four substantive sections of the Clayton Act targeted certain specific types of business conduct. Specifically, Section 2 forbade price discrimination, although it provided for three defenses that have turned out to be fairly large loopholes. Section 3 forbade *tying contracts,* which force a buyer to purchase one good in order to buy another good, and *exclusive dealing arrangements,* which require a buyer to purchase its entire supply of some good from one seller. Section 7 targeted the horizontal mergers that took place after the passage of the Sherman Act, making it illegal for a company to acquire the stock of another corporation if the acquisition would "restrain commerce in any section or community or tend to create a monopoly of any line of commerce." Finally, Section 8 prohibited interlocking directorates between any two competitors, making it illegal for the same person to serve on the board of directors of rival companies.

The Federal Trade Commission (FTC) Act established the Federal Trade Commission, an independent antitrust agency empowered to enforce the Clayton Act. In addition, Section 5 of the FTC Act directs the commission to "prevent persons, partnerships, or corporations, except banks, and common carriers subject to the acts to regulate commerce, from using unfair methods of competition in commerce." Although this seems like a broad mandate, in fact, the FTC had little power in its early days, with its only remedy being the power to issue cease-and-desist orders. Also, economists have pointed out that what some firm regards as an "unfair" method of competition may increase overall competition and benefit consumers.

The next major piece of antitrust legislation in the United States, a product of the Great Depression when countless numbers of businesses failed, is the most controversial. The Robinson-Patman Act concerns price discrimination, amending part of the Clayton Act and adding provisions declaring specific types of price discrimination illegal regardless of the reason underlying their use. In practice, the Robinson-Patman Act often has been used to protect individual firms rather than competition.

In 1950, Congress passed the final major amendment to the Clayton Act. The Celler-Kefauver Act closed a loophole in the original act. Firms had discovered that, as written, the act forbade corporations from merging by buying the stock of their competitors but *not* by directly acquiring rivals' assets. The 1950 amendment put some teeth into the restrictions against anticompetitive mergers.

More than a century of experience with antitrust law has shown the importance of interpretation by the courts. During the past 40 years, antitrust policy itself has undergone a considerable transformation as industrial organization economics has come to play an increasingly important role. This process continues with the outcome unclear. However, it is exciting to know that economic analysis is now central to enforcement policy in the Justice Department and the Federal Trade Commission and to decisions made in courts. The interaction between advances in economic theory and implementation and interpretation of antitrust policy is rewarding for economists and policy makers alike.

## References and Further Readings


Brock, J. W. (Ed.). (2009). *The structure of American industry* (12th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Chamberlain, E. H. (1950). *The theory of monopolistic competition.* Cambridge, MA: Harvard University Press.

Dixit, A., & Nalebuff, B. (1991). *Thinking strategically: The competitive edge in business, politics, and everyday life.* New York: W. W. Norton.

Dixit, A., Reiley, D. H., Jr., & Skeath, S. (2009). *Games of strategy* (3rd ed.). New York: W. W. Norton.

Kwoka, J. E., Jr., & White, L. J. (2009). *The antitrust revolution: Economics, competition, and policy* (5th ed.). New York: Oxford University Press.

Nicholson, W., & Snyder, C. (2007). *Intermediate microeconomics and its application* (10th ed.). Florence, KY: South-Western (Cengage Learning).

Perloff, J. M. (2008). *Microeconomics: Theory and applications with calculus.* Upper Saddle River, NJ: Prentice Hall.

Pindyck, R., & Rubinfeld, D. (2009). *Microeconomics* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Posner, R. A. (1975). The social costs of monopoly and regulation. *Journal of Political Economy, 83,* 807–827.

Tremblay, V. J., & Tremblay, C. H. (Eds.). (2007). *Industry and firm studies* (4th ed.). Armonk, NY: M. E. Sharpe.

Waldman, D. E. (2009). *Microeconomics* (a .learn eBook) (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Waldman, D. E., & Jensen, E. J. (2007). *Industrial organization: Theory and practice* (3rd ed.). Boston: Pearson Addison Wesley.

# 13

# Predatory Pricing and Strategic Entry Barriers

Bradley P. Kamp
*University of South Florida*

The classic definition of predatory pricing is pricing below cost with the intention of running a competitor out of business. In more general terms, predatory pricing is a price reduction that is only profitable because of added market power the predator gains from eliminating, disciplining, or otherwise inhibiting the competitive conduct of a rival or potential rival (Bolton, Brodley, & Riordan, 2000). Claims of large companies preying on their smaller competitors are commonplace, starting during the formation of trusts during the late nineteenth century up through charges against Wal-Mart today. Yet in the two most recent, precedent-setting predatory pricing cases, the Supreme Court observed that "there is a consensus among commentators that predatory pricing schemes are rarely tried, and even more rarely successful" (*Matsushita Electric Industrial Co. v. Zenith Radio Corp.,* 1986), and "predatory pricing schemes are implausible . . . and even more implausible when they require coordinated effort among several firms" (*Brooke Group Ltd. v. Brown and Williamson Tobacco Corp.,* 1993). The goal of this chapter is to present a brief history of predatory pricing, beginning with the era of trusts. The debate as to the rationality of predatory pricing will be examined, followed by a discussion of what is currently recognized by the courts as an act of predation. A look at research since *Brooke Group* ensues. As the recoupment of lost profits during the period of predation is now a necessary condition to be guilty of predatory pricing, strategic entry barriers will be examined. Possible explanations of the prevalence of predatory pricing cases are offered, with a discussion of claims against Wal-Mart to follow. Some short remarks conclude.

## Predation and Trusts

During the era of trust building, predatory pricing was believed to be a commonly employed business practice. This was predation of the price discrimination variety—firms would cut prices in some markets to influence rivals while leaving prices higher in other markets. Many cases have been examined by antitrust scholars, including trusts in gunpowder (Elzinga, 1970), sugar (Zerbe, 1969), cornstarch/syrup (*United States v. Corn Products Refining Co.,* 1916), tobacco (Burns, 1986), and matches in Canada and the British "fighting ships" (Yamey, 1972). Perhaps the best-known accusation was put forth by Ida Tarbell in her 1904 book *The History of Standard Oil Company.* Coupled with testimony during the antitrust case against Standard (*Standard Oil Co. of New Jersey v. United States,* 1911), it was widely believed that Standard did employ predatory pricing along with other anticompetitive practices. Even today, the Standard Oil case is often the first antitrust case presented in introductory economics texts.

Standard was accused of using its monopoly power to price discriminate, undercutting rivals in some locations, while subsidizing losses with profits from other locations. Similarly, it maintained its monopoly position by selectively cutting prices in markets where competitors dared to

enter. It was widely agreed that predatory pricing played an important role in Standard's monopolization of oil refining. So concerned were regulators about the effectiveness of predatory pricing in securing monopoly power that predatory pricing was included in the list of business practices specifically outlawed by the Clayton Act of 1914 and again in the Robinson-Patman Act of 1936.

The Standard Oil case also led to the first major questioning of predatory pricing as a viable business practice. In his groundbreaking paper, John McGee (1958) scoured the court records, closely examining each of Standard's 123 known refinery purchases. He found numerous contradictions to what would be considered classic predation. First, predatory pricing assumes that a large firm locally cuts prices while subsidizing the losses from profits elsewhere (i.e., the deep-pockets argument), yet Standard had less than 10% of the refining market as late as 1870. Predatory pricing strategies cannot explain how Standard obtained the necessary monopoly power to prey. McGee was the first to point out that when engaged in a price war, the larger supplier suffers the larger loss. In many regions, Standard had 75% or more of the local market, meaning that in a below-cost price war, Standard's losses would be at least three times those of the other competitors combined. In addition, a refinery does not just disappear when production stops. There is no reason why a competitor cannot shut down during the price war, then commence production once Standard raises price to recover the losses. McGee posited that obtaining competitors through mergers would be less costly, because the combined firm could earn monopoly profits right away, rather than suffer losses during a price war of unknown length. In addition, there should be some agreeable price, between the rents earned in a competitive market and those earned in a monopoly. McGee found that Standard obtained its monopoly through mergers rather than predation. Most of these mergers were of the friendly type, where owners and managers of the purchased refineries kept their positions. In other cases, owners who had sold to Standard opened new refineries in markets where Standard had a presence. Neither of these should occur if Standard used predation, because such an adversarial relationship would not lead to a cordial working agreement. McGee also refuted the argument that Standard used predation to lower the purchase price. To quote McGee: "I can not find a single instance in which Standard used predatory price cutting to force a rival refiner to sell out, to reduce asset values for purchase, or to drive a competitor out of business" (p. 157).

## Chicago and Post-Chicago Schools

Located at the University of Chicago during his Standard Oil study, McGee is a leading member of the Chicago school of antitrust scholars. The terms *Chicago* and *post-Chicago* arise frequently in the antitrust literature. Most Chicago scholars have some relation to the University of Chicago, but scholars who are skeptical about predatory pricing, regardless of affiliation, are classified as "Chicago school." *Post-Chicago* refers to scholars active primarily in theoretical industrial organization who examine the possibility of predation as an equilibrium strategy. There is no general consensus among post-Chicago scholars as to the viability and frequency of predatory pricing.

The general view of predation held by the Chicago school is that it is not a rational, long-run profit-maximizing strategy. The costs of predation must be offset by future monopoly profits, which are highly uncertain. Without postpredation barriers to entry, the monopoly price cannot be sustained, and no harm comes to consumers. Under this scenario, no action should be taken against aggressive pricing campaigns, even if competitors are injured. The benefits to consumers of vigorous price competition dominate any costs to the firms. Other classic presentations of the Chicago perspective on predatory pricing include Robert H. Bork (1978), Frank H. Easterbrook (1981), and John Lott, Jr. (1999). The current requirement of recoupment crystallized in *Brooke Group* (to be discussed in detail below) was first explicitly defined in a lower court case heard by Easterbrook (*A. A. Poultry Farms, Inc. v. Rose Acre Farms, Inc.,* 1989).

One perceived weakness of McGee's (1958) argument was its somewhat static approach. The effect of predation in one market on rivals or potential rivals in other future markets was never explicitly examined. Game theory allowed economists to explicitly model such strategic behavior. The seminal paper in this area, Selten's (1978) chain store paradox, showed that preying could not be an equilibrium strategy, even when considering its effect on future rivals. A series of papers followed, each relaxing one or more of Selten's assumptions. The authors working in this area form the "post-Chicago school."

The post-Chicago school examines the possibility of successful predation in a formal, equilibrium setting. These game-theoretic models include predation as a possible strategy and find that under certain assumptions, predatory pricing can be a rational, equilibrium strategy. These models expand the finite, full-information models of McGee and Selten into a more complex, dynamic world of imperfect and asymmetric information. The post-Chicago literature analyzes two major sets of models: those based on asymmetric financial constraints and those based on signaling. Ordover and Saloner (1989) present a thorough discussion of the post-Chicago literature.

Models involving asymmetric financial constraints trace their origin to Telser's (1966) "deep-pockets" argument. The predator has greater resources and can outlast the prey during a below-cost price war. In Telser's model, predatory pricing was not part of an equilibrium because under common knowledge, the potential prey would leave at the first opportunity rather than fight a price war it knew it was destined to lose. Predatory pricing does arise as an equilibrium strategy in works by Fudenberg and Tirole (1985, 1986). In these models, there is an information asymmetry in the credit

market. By preying in one period, the predator can lower the expected value of the prey's assets, leaving the prey unable to obtain financing in future periods.

Predatory pricing can be used as a signal to influence expectations of future profitability held by rivals, potential entrants, or even the creditors of the prey. Early papers by Milgrom and Roberts (1982) and Kreps and Wilson (1982) added an information asymmetry to the chain store model. Both showed that if the predator knew more than the potential entrants, predation could be an equilibrium strategy, although not necessarily the only equilibrium. A predatory price might be used to signal low costs (Milgrom & Roberts, 1982) or that the predator is aggressive and prefers preying to accommodating (Kreps & Wilson, 1982). A predatory price might be part of an equilibrium where the exit of a rival does not occur. Predation is used to change the prey's behavior, possibly to obtain better merger terms (Saloner, 1987). Critical to all of the post-Chicago school models is the assumption of an information asymmetry. Predation is successful only when the predator knows something that the prey or financial market does not and can use pricing strategies to influence beliefs about the unknown. When the uncertainty is in both directions (i.e., when the predator does not know everything about the prey as well), the likelihood of successful predation drops.

## What Constitutes Predation

Under the Sherman and Clayton Acts, a firm could be found guilty of predatory pricing if it priced sufficiently lower in markets where it faced competition than in markets where it did not. Predatory pricing cases were infrequent until passage of the Robinson-Patman Act in 1936. Undercutting the price of a local firm by a large, multimarket firm was prima facie outlawed. In the 1940s, the Federal Trade Commission (FTC) stepped up its enforcement of the Robinson-Patman Act, with plaintiffs winning most litigated cases, including at least some cases where there was not clear predation (Koller, 1971). Antitrust scholars refer to this as the "populist era" of predatory pricing enforcement (Bolton et al., 2000).

The Robinson-Patman Act protected businesses from larger rivals. However, little notice was given to the benefits to consumers arising from lower prices and more vigorous competition. A large firm might undercut a local firm's price not because it is a predator but perhaps to capture some of the local firm's rents. The way that predatory pricing was viewed under Robinson-Patman varied considerably. In the precedent-setting 1967 *Utah Pie* ruling, evidence of predatory intent and an unreasonably low price were sufficient for a verdict of predation. However, the *Utah Pie* ruling neglected to specifically define what was meant by an "unreasonable" price. In their 1975 paper, Areeda and Turner tied the definition of a predatory price directly to the costs of the predator. A price was found to be predatory if it was below the predator's average variable cost (used as a proxy for marginal cost). The "Areeda-Turner rule," or some slight variation, was quickly and widely adopted by the courts. Predation suddenly became more difficult to prove: In the 7 years following the article's publication, plaintiffs' success rate dropped to only 8%, as compared to 77% during the populist era (Bolton et al., 2000). To date, the Areeda-Turner rule remains an important factor in determining predatory pricing violations. However, pricing below average variable cost is no longer sufficient to be guilty of predatory pricing.

In the *Matsushita* case, the Supreme Court, for the first time, looked beyond short-run losses and weighed the possibility of recoupment in its decision. The Court held, "The success of any predatory scheme depends upon maintaining monopoly power for long enough to both recoup the predator's losses and to harvest some additional gain" (*Matsushita Electric Industrial Co. v. Zenith Radio Corp.,* 1986). As the plaintiffs failed to offer convincing evidence of possible recoupment, the Court found it unnecessary to conduct a detailed inquiry into whether the defendant's prices were below costs. In the most important precedent-setting case of late, *Brooke Group Ltd. v. Brown and Williamson Tobacco Corp.* (1993), the prospect of recoupment was the primary determinant of whether damage to competition had occurred. At the appellate level, the Fourth Circuit Court of Appeals acknowledged that prices below some measure of costs are necessary to prove predation. However, the court ruled that predation also required "the rational expectation of later realizing monopoly profits. The failure to show this additional aspect is fatal" (*Liggett Group, Inc. v. Brown and Williamson Tobacco Corp.,* 1992). The Supreme Court declined to enter the conflict over measures of costs but concentrated instead on the plausibility of recoupment. The Court ruled that when assessing the plausibility of recoupment, the analysis must first "demonstrate that there is a likelihood that the predatory scheme alleged would cause a rise in prices above the competitive level" and these prices during the recoupment stage must "be sufficient to compensate for the amounts expended on the predation, including the time value of the money invested in it" (*Brooke Group Ltd. v. Brown and Williamson Tobacco Corp.,* 1993, p. 2589). The Court offered examples of where predation would be unlikely to lead to recoupment: where barriers to entry are low or where the market structure is diffuse and competitive. In the final ruling on *Brooke Group,* the Court found that recoupment, and therefore predatory pricing, was implausible given the oligopolistic nature of the market. For a thorough examination of recoupment standard put forth in *Brooke Group,* see Elzinga and Mills (1994).

To be found guilty of predatory pricing today, evidence must be presented that the defendant priced below some measure of cost *and* that recoupment of losses suffered during the period of predation (recoupment of losses is the subject, not pricing below cost and recoupment) is probable. The recoupment standard is critical and difficult for the plaintiff to prove. Zerbe and Mumford (1996) provide a

detailed examination of the first 14 post–*Brooke Group* cases. In fact, since *Brooke Group,* the FTC has not successfully prosecuted a single firm.

## Current Research in Predatory Pricing and Strategic Entry Barriers

While the courts of late remain skeptical about the prevalence of and damages caused by predatory pricing, it continues to be an area of active research. The requirement that price lies below some measure of cost raises some concerns. Edlin (2002) examines the possibility of predation when prices stay above costs. In addition, cost-based rules would miss an important method of predation—the raising of competitor's costs. Granitz and Klein (1996) applied econometric techniques not available to McGee and presented evidence that Standard Oil did in fact use this method of predation.

Theorists working in the field of predatory pricing express concern that the courts have rarely cited the strategic (i.e., post-Chicago) literature, even though this literature offers an array of models where predation is a rational, profit-maximizing strategy. It is important to note that in every strategic model, predation occurs in equilibrium only if recoupment is possible. Of course, recoupment is only possible if there is some barrier to entry, so this line of research could be thought of as the study of strategic entry barriers. The arguments put forth in the strategic entry barrier literature are not in conflict with the crux of the *Brooke Group* ruling. Rather, some economists worry that the courts are inadequately assessing the probability of recoupment. Courts that fail to recognize market imperfections might miss possible ways these imperfections can be exploited to keep rivals out.

Critical in McGee's (1958) critique of the Standard Oil case was the assumption that costs were similar across firms. When markets work well and technology is easily transferable, such as in oil refining, this assumption is reasonable. But some markets fail to meet these conditions. One example is when production exhibits a learning curve, where costs fall as a producer obtains more experience in production. Cabral and Riordan (1994, 1997) show that a learning curve can be manipulated to create a barrier to entry. In such models, a firm preys early to increase output, thereby moving farther down the learning curve and eventually gaining an entry-deterring cost advantage. This behavior clearly meets the predatory pricing standards put forth in *Brooke Group.* However, the welfare effects of such predation are indeterminate. As stated in the conclusion of the 1997 paper, "The information requirements of fashioning an effective legal rule against harmful predation are formidable" (p. 168). A similar stance was taken by Chiaravutthi (2007), who studied the possibility of successful predation in a market with network externalities. A network externality occurs when the benefit of joining a network increases with the number of members. Chiaravutthi found that in a laboratory setting, predatory pricing is often a successful strategy. Predation increases membership early, raising the value to potential members above that of other networks and creating an entry barrier. The welfare effects of successful predation were again ambiguous, and the author concluded that "all in all, it seems extremely difficult to frame a legal rule that would make the correct diagnosis in all cases" (p. 169). Note that it was the market failure, not the laboratory setting, that resulted in successful equilibrium predation in Chiaravutthi's model. Earlier lab testing of predatory pricing in well-behaved markets was not quite as successful (see Gomez, Goeree, & Holt, 1999, for a survey of laboratory-based predatory pricing research).

Since Chamberlain (1933), it has been widely accepted in the industrial organization literature that product differentiation is a potential source of market failure. Schmalensee (1978) showed that product differentiation can be an effective barrier to entry. Lindsey and West (2003) examined the effects of predatory pricing in a market with differentiated products. They found (via simulation) that with product differentiation, predatory pricing is a possible equilibrium strategy. However, the authors did not explicitly address recoupment or welfare effects.

During the 1980s, much research was undertaken in finance on the imperfections inherent is capital markets. Bolton and Scharfstein (1990) showed that these imperfections can lead to an entry barrier necessary for successful predation. Entrants into an industry typically have implicit agreements with their backers for additional postentry financing. This financing is often conditional on early post-entry performance. By preying, an incumbent firm can reduce the entrant's profits, increasing the chance that additional funding will be denied. Additional work in the field of financial market predation includes Bulow and Rogoff (1989) and Diamond (1989). All of the authors stress the relevance of their models in rapidly evolving industries where firms rely heavily on venture capital for financing.

The discussion on recoupment standards continues. Iacobucci (2006) examines Canada's wavering on recoupment post *Brooke Group.* He presents a strong argument that recoupment tests reduce the chance of improperly characterizing lawful price reductions as predatory. A debate between some of the brightest scholars working in the field was published in the *Georgetown Law Journal* in 2001 (Bolton, Brodley, & Riordan, 2000, 2001; Elzinga & Mills, 2001). Bolton et al. (2000, 2001) take the view that the courts need to pay more heed to strategic models. They present a series of tests that would better enable the courts to discover possible strategic entry barriers that currently go unnoticed. Elzinga and Mills (2001) point out the strong assumptions in the strategic models and worry that applying theoretical models to actual firm behavior would be extremely difficult. Perhaps their position is put forth best in the following quote: "If you are hunting for a predator and mistakenly shoot a competitor, you injure consumers" (p. 3).

## The Prevalence of Predatory Pricing Cases

If predatory pricing is not a viable business practice, why do we still see lawsuits claiming predation? The first reason was discussed above: Strategic predatory pricing is taking place, but the courts are not sophisticated enough to recognize it. Perhaps a plaintiff knows competition has been harmed by predation and that recoupment will be possible. Given that antitrust awards are automatically trebled, it is worth taking the chance to conceivably become the new precedent-setting case.

Baumol and Ordover (1985) provide another explanation of predatory pricing cases: use of the antitrust laws to facilitate tacit collusion. Collusion between competitors to restrict output (i.e., raise price) is against U.S. law. However, it is widely held that there is extensive tacit or informal collusion throughout the economy. Baumol and Ordover present a scenario where industry rivals threaten predatory pricing suits against competitors that make a competitive price cut. Given the expense of fighting a predatory pricing lawsuit (e.g., American Airlines spent $20–$30 million to fight the suit brought by Continental; Clouatre, 1995), such threats can be effective deterrents to competition. Harrington (2004) formally examined the effects of antitrust laws on the stability of cartels. He found that it is possible that such laws, which were enacted to fight cartels, might actually increase the viability of a cartel.

A simpler, nonstrategic explanation is that the plaintiff does not know its rival's costs. What seems to be a predatory price (i.e., a price significantly below the plaintiff's costs) might not actually be below the defendant's costs. Kamp and Thomas (1997) modeled such a situation, where a rival firm's quality (and therefore costs) was uncertain. Due to a secret quality cut, what appeared to be a predatory, below-cost price was actually a price that was still above cost. The necessary condition of a below-cost price is not met, and the courts would find in favor of the defendant. A final possibility is that a below-cost price is set to capture a larger market share, without regard to possible recoupment. Thomas and Kamp (2006) present several plausible reasons why at least some managers, and perhaps even many managers, sometimes choose to pursue higher market share by cutting prices, even when the price cuts and higher market share reduce profitability. Under either of these types of predation, current predatory pricing policy yields desirable antitrust enforcement outcomes.

## Wal-Mart

Since 1990, no firm in the United States has generated more claims of predatory pricing than Wal-Mart. The Bentonville, Arkansas–based retailer obtained its position as the largest retailer in the United States primarily by offering "everyday low prices." Sometimes its competitors believe that these low prices are too low, even below costs. Perhaps the most important predatory pricing suit against Wal-Mart was filed in the Faulkner (Arkansas) County Chancery Court by three local pharmacies in 1991 (*American Drugs, Inc. v. Wal-Mart Stores, Inc.,* 1993). The plaintiffs claimed that Wal-Mart violated the 1937 Arkansas Unfair Trade Practices Act by routinely selling pharmaceutical products below cost. Strong evidence was provided that Wal-Mart priced certain items below its wholesale cost. One of the plaintiffs testified that he stocked his own store from Wal-Mart, as the retail price was below the lowest wholesale price he could find (Kurtz, Keller, Landry, & Lynch, 1995). This strategy is not unique to the Wal-Mart case: During an early twentieth-century attempt at predatory pricing, Dow purchased bromine offered on the U.S. market at below costs by its German competitor and resold it for a profit in Europe (Levenstein, 1997). In 1993, the Faulkner Court found Wal-Mart guilty of predatory pricing. Wal-Mart was ordered to stop charging below-invoice prices for its pharmaceuticals and to pay the three plaintiffs $289,407 (Boudreaux, 1996). Wal-Mart appealed to the Arkansas Supreme Court (*Wal-Mart Stores, Inc. v. American Drugs, Inc.,* 1995), and in January 1995, the Court ruled by a 4–3 decision to overturn the lower court verdict and dismiss the case.

More recently, in 2001, Wal-Mart and the Wisconsin Department of Agriculture reached a settlement over claims that Wal-Mart sold butter, milk, laundry detergent, and other staples below costs in five Wisconsin cities. In the settlement, Wal-Mart admitted no wrongdoing and was not required to pay a fine. However, it agreed to pay double or triple fines for any future violations. In 2007, Wal-Mart was forced to eliminate its low-priced generic prescription plan for certain drugs in nine states, as their price was below their true costs. While questions remain as to whether Wal-Mart's aggressive pricing harms competition, there is no debate as to the benefits of low prices to its customers.

## Conclusion

The debate about the existence and feasibility of predatory pricing shows no sign of stopping. Currently, the courts take a somewhat skeptical view of the viability of predatory pricing. Showing that price was below cost is not enough: It is now the plaintiff's responsibility to show that the predator will be able to recover its losses suffered while preying. Much to the chagrin of theorists working in industrial organization, the courts of late have been unwilling to refer to the strategic literature when ruling. The makeup of courts changes, however, and there is no guarantee that we have reached a steady state regarding predatory pricing.

As long as rival firms engage in intense price competition, there will continue to be claims of predation. And as long as firms price below costs, economists will attempt to model predatory pricing as an equilibrium, profit-maximizing strategy.

## References and Further Readings


*A. A. Poultry Farms, Inc. v. Rose Acre Farms, Inc.,* 881 F.2d 1396, 7th Circuit (1989).

*American Drugs, Inc. v. Wal-Mart Stores, Inc.,* E-92–1158 (1993).

Areeda, P., & Turner, D. (1975). Predatory pricing and related practices under Section 2 of the Sherman Act. *Harvard Law Review, 88,* 697–733.

Baumol, W., & Ordover, J. (1985). Use of antitrust to subvert competition. *Journal of Law and Economics, 28,* 247–265.

Bolton, P., Brodley, J., & Riordan, M. (2000). Predatory pricing: Strategic theory and legal policy. *Georgetown Law Journal, 88,* 2239–2330.

Bolton, P., Brodley, J., & Riordan, M. (2001). Predatory pricing: Response to critique and further elaboration. *Georgetown Law Journal, 89,* 2495–2529.

Bolton, P., & Scharfstein, D. (1990). A theory of predation based on agency problems in financial contracting. *American Economic Review, 80,* 93–106.

Bork, R. H. (1978). *The antitrust paradox: A policy at war with itself.* New York: Free Press.

Boudreaux, D. (1996). Predatory pricing in the retail trade: The Wal-Mart case. In M. Coate & A. Kleit (Eds.), *The economics of the antitrust process* (pp. 195-216). Boston: Kluwer.

Brooke Group Ltd. v. Brown and Williamson Tobacco Corp., 113 U.S. 2578 (1993).

Bulow, J., & Rogoff, K. (1989). A constant recontracting model of sovereign debt. *Journal of Political Economy, 97,* 155–178.

Burns, M. (1986). Predatory pricing and the acquisition costs of competitors. *Journal of Political Economy, 94,* 266–296.

Cabral, L., & Riordan, M. (1994). The learning curve, market dominance, and predatory pricing. *Econometrica, 62,* 1115–1140.

Cabral, L., & Riordan, M. (1997). The learning curve, predation, antitrust and welfare. *Journal of Industrial Economics, 45,* 155–169.

Chiaravutthi, Y. (2007). Predatory pricing with the existence of network externalities in the laboratory. *Information Economics and Policy, 19,* 151–170.

Chamberlain, E. (1933). *The theory of monopolistic competition.* Cambridge, MA: Harvard University Press.

Clouatre, M. (1995). The legacy of Continental Airlines v. American Airlines: A re-evaluation of predatory pricing theory in the airline industry. *Journal of Air Law and Commerce, 60,* 869–915.

Diamond, D. (1989). Reputation acquisition in debt markets. *Journal of Political Economy, 97,* 829–862.

Easterbrook, F. (1981). Predatory strategies and counterstrategies. *University of Chicago Law Review, 48,* 263–337.

Edlin, A. (2002). Stopping above-cost predatory pricing. *Yale Law Journal, 111,* 941–991.

Elzinga, K. (1970). Predatory pricing: The case of the gunpowder trust. *Journal of Law and Economics, 13,* 223–240.

Elzinga, K., & Mills, D. (1994). Trumping the Areeda-Turner test: The recoupment standard in *Brooke Group. Antitrust Law Journal, 62,* 559–584.

Elzinga, K., & Mills, D. (2001). Colloquy: Predatory pricing and strategic theory. *Georgetown Law Journal, 89,* 2475–2494.

Fudenberg, D., & Tirole, J. (1985). *Predation without reputation* (Working Paper 377). Cambridge: MIT.

Fudenberg, D., & Tirole, J. (1986). A signal-jamming theory of predation. *RAND Journal of Economics, 17,* 366–376.

Gomez, R., Goeree, J., & Holt, C. (1999). *Predatory pricing: Rare like a unicorn* (Working paper). [Charlottesville, VA]: University of Virginia.

Granitz, E., & Klein, B. (1996). Raising rivals' costs. *Journal of Law and Economics, 39,* 1–47.

Harrington, J. (2004). Cartel pricing dynamics in the presence of an antitrust authority. *RAND Journal of Economics, 35,* 651–673.

Iacobucci, E. (2006). Predatory pricing, the theory of the firm, and the recoupment test: An examination of recent developments in Canadian predatory pricing laws. *Antitrust Bulletin, 51,* 281–323.

Kamp, B., & Thomas, C. (1997). Faux predation in markets with imperfect information on product quality. *Southern Economic Journal, 64,* 555–566.

Koller, R. (1971). The myth of predatory pricing. *Antitrust Law and Economics Review, 4,* 105–123.

Kreps, D., & Wilson, R. (1982). Reputation and imperfect information. *Journal of Economic Theory, 27,* 253–279.

Kurtz, D., Keller, S., Landry, M., & Lynch, D. (1995). Wal-Mart fights the battle of Conway. *Business Horizons, 38,* 46–50.

Lindsey, R., & West, D. (2003). Predatory pricing in differentiated products retail markets. *International Journal of Industrial Organization, 21,* 551–592.

Levenstein, M. (1997). Price wars and the stability of collusion: A study of the pre–World War I bromine industry. *Journal of Industrial Economics, 45,* 117–137.

*Liggett Group, Inc. v. Brown and Williamson Tobacco Corp.,* 964 F. 2d 335, 4th Circuit (1992).

Lott, J. R., Jr. (1999). *Are predatory commitments credible? Who should the courts believe?* Chicago: University of Chicago Press.

*Matsushita Electric Industrial Co. v. Zenith Radio Corp.,* 475 U.S. 574, 588–89 (1986).

McGee, J. (1958). Predatory price cutting: The Standard Oil (N.J.) case. *Journal of Law and Economics, 1,* 137–169.

Milgrom, P., & Roberts, J. (1982). Predation, reputation and entry deterrence. *Journal of Economic Theory, 27,* 280–312.

Ordover, J., & Saloner, G. (1989). Predation, monopolization and antitrust. In R. Schmalensee & R. Willig (Eds.), *Handbook of industrial organization* (Vol. 1, pp. 537–596). Amsterdam: North-Holland.

Saloner, G. (1987). Predation, merger and incomplete information. *RAND Journal of Economics, 18,* 165–186.

Schmalensee, R. (1978). Entry deterrence in the ready-to-eat breakfast cereal industry. *Bell Journal of Economics, 9,* 305–327.

Selten, R. (1978). The chain-store paradox. *Theory and Decision, 9,* 127–159.

*Standard Oil Co. of New Jersey v. United States,* 221 U.S. 1 (1911).

Tarbell, I. (1904). *The history of Standard Oil Company.* New York: McClure Phillips.

Telser, L. (1966). Cutthroat competition and the long purse. *Journal of Law and Economics, 9,* 259–277.

Thomas, C., & Kamp, B. (2006). Buying market share: Agency problem or predatory pricing? *Review of Law and Economics, 2,* 1–24.

*United States v. Corn Products Refining Co., et al.,* 234 F. 964 (1916).

*Utah Pie Co. v. Continental Baking Co.,* 386 U.S. 685 (1967).

*Wal-Mart Stores, Inc. v. American Drugs, Inc.,* 819 S.W. 2d 30 (1995).

Yamey, B. (1972). Predatory price cutting: Notes and comments. *Journal of Law and Economics, 15,* 129–142.

Zerbe, R. (1969). The American Sugar Refining Company, 1887–1914: The story of a monopoly. *Journal of Law and Economics, 12,* 339–375.

Zerbe, R., & Mumford, M. (1996). Does predatory pricing exist? Economic theory and the courts after *Brooke Group. Antitrust Bulletin, 41,* 949–985.

# 14

# Labor Markets

Kathryn Nantz
*Fairfield University*

Work is a part of life for almost all people around the world. Though the types of work people do and the conditions under which that work is done vary endlessly, people get up each morning and choose to use their human capital in ways that generate some sort of productive good or service or that help prepare them to be productive economic citizens in the future. Some of this work is done in the privacy of the home, where beds are made, children are raised, and lawns are mowed. While this unpaid productive activity is essential to a well-functioning economy, this chapter addresses work time and skills that are sold in markets in exchange for wages and other compensation.

The purpose of this chapter is to explore the unique nature of labor markets and to consider how these markets will evolve in response to changes in the nature of the work people do over time. Use of labor, like any economic resource, has to be considered carefully in light of productivity and opportunity costs. Though many factors affect this decision-making process, in most cases labor is allocated by market forces that determine wages and employment.

That said, several characteristics of labor as a resource create complexities. First, the demand for any type of labor services is derived from, or dependent on, the demand for the final product that it is used to produce. This means that highly trained and productive workers are only as important in production in an economy as there is demand for the product they produce. Second, because labor cannot be separated from the particular persons who deliver the resource, the scope of responses to labor market decisions is broad and affects outcomes in significant ways. The sale of labor services generates the majority of household income around the world—income that is used to sustain workers and their families. This means that labor markets determine, to a large extent, what resources a household has available and thus the quality of life for the members of the household. Decisions become very complex when workers and their families begin considering not only job market choices but also premarket education and training that might be required to prepare individuals for particular occupations.

## Theory of Labor Market Allocation

As with all markets, buyers (employers or contractors) and sellers (employees) communicate their needs and offers with one another and exchange labor services for wages and other compensation. Some important assumptions underlie this model. First, assume that the wage and other monetary compensation is the most important determinant of behavior. This allows the construction of a model that has wages on the $y$-axis and quantity exchanged on the $x$-axis. Everything else held fixed, buyers and sellers will respond most directly to price signals exchanged between the two groups. Second, assume that workers are somewhat homogeneous—that is, that workers can be easily substituted for one another in any particular market. (Differences across workers will be explored later in this chapter.) Third, assume that workers are mobile and can move to places where there is excess demand for labor with their particular skills and away from places where there is excess supply of labor with their particular skills. Finally, assume that wages are flexible; they can move up or down in response to market signals.

Given these assumptions, the next step is to consider the behavior of buyers and sellers in what can be referred to as perfectly competitive markets for labor. In such markets, many relatively small employers hire relatively small amounts of labor; neither employers nor employees acting alone are a significantly large enough share of the market to be able to affect market wages. An example of this might be a college in New York City that hires administrative assistants. There are many such assistants in the market looking for jobs and many alternative employers, so both buyers and sellers have to accept the going wage for such work.

## Labor Demand

Employers seek workers based on workplace needs and based on the demand for the good or service being produced. In addition to the market wage, employers consider the productivity of labor, the ability to substitute across other inputs in the production process, and the prices of other inputs when making hiring decisions.

Labor productivity is determined by a variety of factors, including human capital investments made by the worker himself or herself or the employer, skills and talents, and the quantity of capital and technology that the worker has at his or her disposal. As might be assumed, the greater the investment made in an individual worker, the greater his or her productivity. Like with any other investment opportunity, the investor spends money now (e.g., gets a master's degree in social work or an MBA), hoping to eventually reap the benefits, in terms of greater income earning potential, in the future.

Substitution of inputs can be easy or difficult, depending on the production process being considered. For example, in an accounting firm, junior and senior partners might be equally productive (though not, perhaps, in the minds of important clients). Junior and senior partners can be relatively easily substituted for one another as tax forms are prepared. However, in a telemarketing firm, each caller needs to have his or her own phone. If an employer hires more callers but buys no more phones, no additional calls will be made. It might be easy to substitute junior for senior account managers, but it is impossible to substitute callers for telephones.

Understanding these sorts of trade-offs is very important for a firm; as the prices of inputs (junior and senior account managers or callers and phones) change, the employer will want to shift the input mix so that output is produced as efficiently and cheaply as possible. However, depending on the degree of substitutability, the changing input prices will create incentives to use more of the input that is becoming relatively cheaper and less of the input that is becoming relatively more expensive.

In spite of these other factors and their importance, firms tend to hire a greater number of hours or workers in the market, and more firms become buyers in a given market, as wages and compensation fall—everything else held constant. This empirically based conclusion is consistent with the law of demand. There is an inverse relationship between the quantity of labor demanded in a market and the price of labor in that market.

## Labor Supply

Workers seek jobs based on their own skills and talents and based on the variety of factors that help to determine their income needs. Preferences over alternative jobs are important—some workers seek jobs that provide them with a sense of pride, accomplishment, and satisfaction. In other cases, workers have a preference for maximizing their own personal income, and so they search for jobs that are high paying, regardless of their other characteristics. Happily, the diversity of jobs combined with the heterogeneity of workers and their preferences generates labor markets that provide incentives for employers and for workers. Jobs that are distasteful to many workers for one reason or another (washing windows on skyscrapers or cleaning out pig sties) attract workers who are less averse to the negative aspects (heights or smells), and/or the work commands wage premiums that accrue from smaller pools of available workers. Thus, workers sort toward jobs that best meet their particular preferences and monetary needs.

Most often, workers consider the expected wage to be a key factor, if not the key factor, in determining what jobs to seek and accept. For some jobs where compensation is based on productivity (sales commissions, as an example, for real estate agents), there can be significant wage uncertainty. This uncertainty means that a job with a high wage might not be as appealing as a job with a lower wage that offers more security. For example, a worker might put her desire to be a professional tennis player aside in favor of the more stable employment position of a bookkeeper if she has elderly parents who need her care. High wages alone are not enough to attract workers to jobs; it is the entire employment package, and the level of employment and compensation risk, that must be considered when choosing a labor market to enter.

Given these other factors and their importance, workers tend to offer more hours in the market, and more workers enter a given market, as wages and compensation rise—everything else held constant. This empirically based conclusion is consistent with the law of supply that governs most market settings. There is a direct relationship between the quantity of labor supplied to a market and the price of labor in that market.

## Market Equilibrium

When economists speak of equilibrium, they are using the term as any other scientist would—as a state of the world that, once reached, will not change unless a significant force acts upon it. Equilibrium price and quantity in a

labor market is reached when there is no further tendency for wages or quantity of labor bought and sold to change, unless there is a change in the market that affects the demand curve or the supply curve.

Consider Figure 14.1. At W2, Q3 workers are demanded, and Q3 workers are supplied. Note that if the wage is W1, the quantity of workers demanded is greater than the quantity of workers supplied. There is a shortage of labor; the quantity of labor employers are willing and able to hire is greater than the quantity of labor people are offering for sale. Wages will rise as firms outbid one another to attract the workers they need. On the other hand, at W3, the quantity of workers demanded is less than the quantity of workers supplied. There is a surplus of labor; the quantity of labor employers are willing and able to hire is less than the quantity of labor people are offering for sale. Wages will fall as workers have to accept lower wages if they hope to gain employment. Thus, the only point at which wages and employment will not be under pressure to rise or fall is at W2, and it is defined as the current equilibrium in this market.

Note several things about this equilibrium price and quantity. First, economists often refer to this point as *market clearing;* this means that everyone who wants to buy labor at this wage rate can and everyone who wants to sell labor at this wage rate can. There is no excess supply or excess demand at the equilibrium wage. Second, although the market clears at this point, it is not necessarily a "good" wage and quantity combination. There are people in the market who would like to purchase labor at wages below W2 who cannot do so (Q3–Q6), and there are people in the market who would like to supply labor at wages above W2 who cannot find jobs (Q3 and above). They are left out of the market under these supply and demand conditions, and that might not be good for the workers or their families if they are unemployed or for the firms if they need workers to produce output to sell in the marketplace. Finally, note that some markets are very volatile, and so even though there are equilibrium wages and quantities exchanged, supply and/or demand are shifting significantly and often. For example, consider dockworkers. When a ship arrives at a port, the workload is heavy and demand for workers is high. However, when there are no ships in port, either because demand for the goods coming in has fallen or because supply has been affected by weather or other factors, demand for workers is low or nonexistent. These sorts of big swings in demand for workers can lead to significant wage volatility. It will become clear that unique sorts of contractual arrangements will be required to be sure that workers will be available for these types of tasks.

## Imperfect Competition

The perfectly competitive market model applies when there are many, many workers in a market that also supports many, many employers. However, what if this is not the case? What if there is a single large employer that dominates the market for a particular type of labor? For example, major league baseball employs almost all of the professional baseball players in the United States. How does this situation impact the distribution of resources in a labor market? How will the equilibrium wage and quantity of labor exchanged be different under these circumstances?

*Monopsony power* refers to the ability of a single employer to control the terms of employment for all of, or a portion of, its workforce. The first significant examples of monopsonies emerged during the industrial revolution when factories (coal mines, steel mills, etc.) opened in small towns; the mine or the factory became the most significant employer in the town or local area, and workers

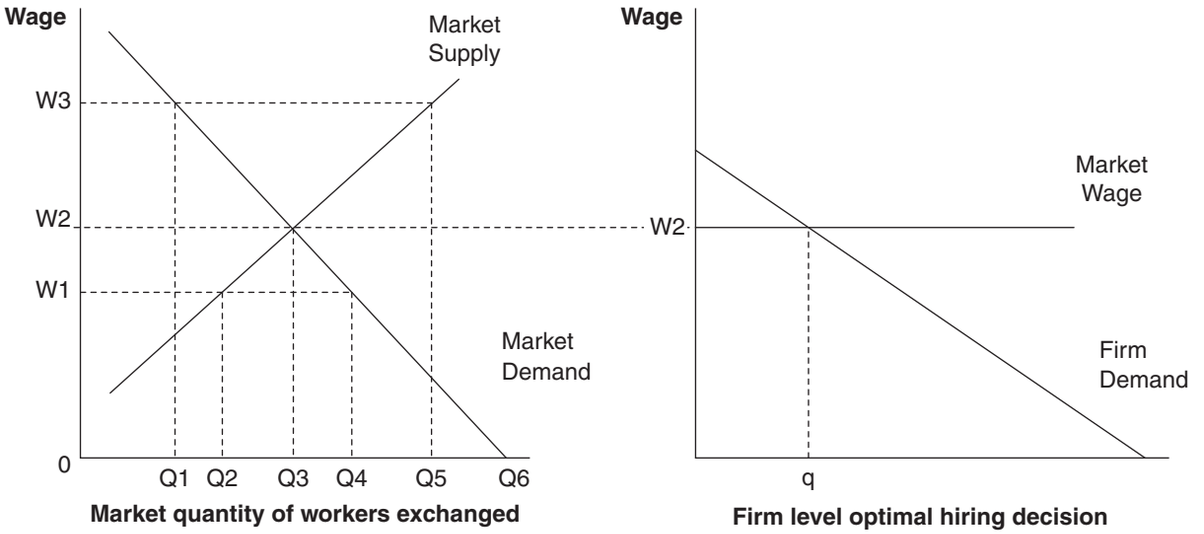


**Figure 14.1** Perfectly Competitive Market for Labor and Firm Level Employment

had few other job opportunities. Because workers had limited alternative employment options, the firms were able to pay lower wages and exploit the workforce in a variety of other ways. In some cases, firms also established company stores that became monopolies in the provision of goods and services. Thus, the workers were subject to monopsony power in the terms of their employment and monopoly power in the markets in which they spent their income.

In response to this exploitation of workers by monopsonistic employers, workers formed unions—collective organizations of sellers that formed to gain some bargaining power in negotiations with their employers over the terms of their contracts. These unions gained legal status in most countries around the world in the late nineteenth and early twentieth centuries. However, in some countries that are at the earlier stages of industrialization, unionization and other types of worker movements may still be illegal or nonexistent. Union history and activity is addressed in a separate chapter in this collection.

Consider the model in Figure 14.2. The labor demand curve, described as for the perfectly competitive market in Figure 14.1, depends on the output market in which the firm sells its final product and the productivity of its workers. However, now rather than paying a wage imposed by the free marketplace, firms have the ability to set wages at the lowest level possible. Firms are not assumed to be wage discriminators; all workers are paid the same amount per quantity supplied. (Wage discrimination is possible but is not considered here.) The key is that firms pay the lowest price they possibly can, given by the market supply curve at the optimal level of employment, and then increase the wage for all by the smallest amount possible, moving up the supply curve, when employment is increased by a small amount.

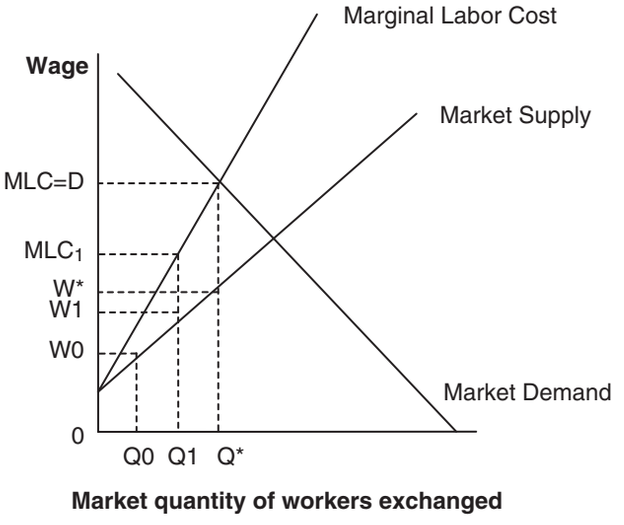


**Figure 14.2** Monopsony Labor Market

Using Figure 14.2, suppose the current level of employment is Q0 and the wage is W0. If the labor monopsonist chooses to increase employment by one worker, labor costs increase by the amount of the wage payment to the new worker but also by the increased wage that has to be paid to all of the existing workers. Firms cannot discriminate and so must pay the same higher wage to all workers of the same type if a marginal worker is to be hired. Thus, the increase in the costs of employment to the firm in this market when a marginal worker is hired is greater than just the wage paid to the new worker; the marginal labor cost (MLC) of hiring one additional unit of labor is *greater than* the wage paid to that one marginal unit. The MLC curve lies everywhere above the supply curve because at any quantity of labor, MLC is greater than the wage.

In these imperfectly competitive markets, firms choose to hire where MLC intersects the demand curve, at MLC = D and Q*. This means the amount of money that comes into the firm when a marginal unit of labor is hired is just equal to the amount of money that goes out when the marginal unit of labor is hired. However, firms are able to pay a wage that is lower than the value of the marginal worker to the firm (W*). This wedge between the value of the worker to the firm and the wage paid has ignited anger among philosophers and workers around the world for hundreds of years. They read Karl Marx's *Communist Manifesto* and believe in the labor theory of value, that the entire value of a final good or service should accrue to the labor inputs that were used to produce it. This alternative to a market model of wage determination had tremendous impacts on economies around the world during the twentieth century and will continue to affect decisions regarding trade-offs between the efficiency and equity of economic outcomes for years to come.

## Applications and Empirical Evidence

There are endless applications of labor market theory to be found in economies around the world. Three of the most significant will be considered here, but the ideas presented can be adapted to fit a variety of other circumstances.

### Economics of the Household or the New Home Economics

Increased labor force participation rates for women around the world, particularly in industrialized countries, were a major feature of the twentieth century. In addition to changing the landscape of labor markets, this phenomenon has had tremendous impacts on households, on families, and on the ways that cultures proscribe the management of unpaid work within the household. According to the U.S. Census Bureau, in 1900, 20% of women in the United States participated in the labor force (U.S. Department of Commerce, 1975). By 2000, that

percentage had increased to 59.9% (International Labour Office [ILO], 2003). By contrast, in 2000 the female labor force participation rate was only 16.3% in Pakistan, 39.6% in Mexico, and 49.2% in Japan. These differences are significant and reflect alternative cultural norms and government support for medical care, child care, and other such programs that make it easier for women to enter the workforce.

In the United States, women began entering the labor market in significantly greater numbers during World War II to replace male employees who were fighting overseas. However, at the end of the war, women were forced out of employment in many cases by men, who were assumed to be heads of households, returning from active duty. It was not until the 1960s that women began to make their way back to work. Many factors can account for this, including the development of safe and reliable birth control, increased access to college and university education for women, declines in birthrates, and increases in divorce rates.

Economists like Gary Becker and Jacob Mincer, who first began investigating family decision making in the early 1960s, used models of international trade to explain how the law of comparative advantage applied equally well to household specialization and trade and to international specialization and trade. These models showed that, when the opportunity costs differed for family members producing the same goods, household production would be more efficient if each member produced those goods and services that they produced with lowest opportunity costs. Trade within the household ensured that each member would be able to consume a mix of goods and services.

This model helped provide some insights into the nature of household production, but it immediately led to alternative explanations and theories. Economists like Barbara Bergman and Julie Nelson have cited alternative explanations for the household structure that we most commonly see around the world—the male head of householder working in the labor market and supporting the efforts of other family members who are engaged in production within the household or human capital formation. One explanation might be found in bargaining models, which describe women and children who are less powerful in the household due to the absence of monetary income or having to depend on an altruistic head of household for economic resources. This means that the typical household structure is the result of differences in access to resources and bargaining power rather than any sort of efficiency in allocation processes or gains from trade explanation. In other cases, the household resource distribution system could be modeled as a Marxist process of exploitation; the "haves" (aka workers in the labor market) exploit the "have nots" (aka nonmarket workers in the household) to pursue their own self-interest and personal resource accumulation. These alternative models of resource allocation within the household have become increasingly important as economists have tried to better understand the nature of social problems like discrimination and domestic abuse.

## Differences in Wages Across Occupations

In most economies, on average, doctors make more money than nurses, highway construction workers make more money than assembly line workers, plumbers make more money than administrative assistants, and stockbrokers make more money than teachers. Why is there so much wage dispersion in labor markets? What causes wages to vary so significantly across different types of workers and occupations?

There are a variety of answers to this question, but the key is that wages are often used to compensate for other aspects of a job that make it more or less attractive. For example, consider two jobs that are very similar in many ways: doctor and nurse. They both work in the same offices or hospitals, they both care for patients who have some form of health issue, and they both report to the same board of trustees or corporate owners of a health care facility. Because of these similarities, one might make an argument that the pay for these two occupations should be equal—what accounts for the differential in wages?

The key here lies in the significant difference in education and training required by doctors compared to nurses, in most cases. According to the Bureau of Labor Statistics, it takes 11 years of higher education, on average, to become a doctor, while the average registered nurse has a 4-year bachelor's degree (U.S. Department of Labor, 2008–2009a, 2008–2009b). If people are to find the additional time spent in education and training, not to mention the added responsibility of being a doctor rather than a nurse, worthwhile, there must be some hope of a future payoff. It is true that many doctors find great satisfaction in helping patients, but additional pay into the future must be expected in order to repay student loans and to cover the opportunity cost of lost wages during long years in school. This sort of return to education and training compensates for an aspect of becoming a doctor that is negative and that would discourage entrants from pursuing the occupation.

In another example, highway construction workers versus assembly line workers, compensation is provided for risk of injury on the job. Though both these occupations typically require dexterity, strength, and concentration, the highway construction worker faces a much greater risk of injury while at work than does the assembly line worker. According to a brief prepared by Timothy Webster (1999) for the Bureau of Labor Statistics, the number of injuries on the job for workers in construction and manufacturing jobs were almost equal; however, construction workers were almost four times as likely to die on the job than manufacturing workers. In this case, the higher wages paid to construction workers, who have similar levels of education and training as manufacturing workers, compensates for the much higher risk of death while on the job.

In other cases, higher wages compensate for unpleasantness on the job. Some of the most unpleasant working conditions can be found in meatpacking plants or in other food processing plants. Steel mills are notoriously hot and loud. Offshore oil drilling requires long periods away from home. Work in chemical plants may increase the risk of illness or cancer. All of these negative aspects of jobs lead to lower supplies of labor at any wage rate and hence higher wages.

One last aspect of an occupation to be considered is wage risk. Though two jobs, stockbroker and college professor, might have the same *expected* wage, the level of risk and uncertainty can be very different. For example, according to the 2005–2006 annual survey by the American Association of University Professors (2006), the average salary for a full professor was $94,738; typically a full professor has job and income security guaranteed by tenure but little hope of additional compensation. The Bureau of Labor Statistics cites the average salary for financial services sales agents in 2006 to be $111, 338; the bonuses and extra compensation available to such workers runs into the millions of dollars in some cases, but as we saw during 2008, job security in such positions is nonexistent when financial services industries feel the sting of recession (U.S. Department of Labor, 2007). Wage and employment risk must be compensated for by the prospect of big wins and large bonus plans.

Given the model of labor markets presented earlier, it is clear that interaction between the demand for labor and the supply of labor determines the equilibrium wage to be paid to workers. The sorts of job characteristics described above all affect the size of the applicant pool available to take on certain types of jobs and hence the position of the labor supply curve.

## Unemployment

Though labor markets clear when quantity demanded is equal to quantity supplied, as described above, there are workers who are still seeking to work once the market clears, just at wages that are above market equilibrium. When people are actively seeking work but cannot find it, they are officially counted among the unemployed. Though in most all markets for goods and services that achieve equilibrium there are buyers and sellers who cannot participate because prices are too high or too low, when this happens in labor markets, households are left without income and other compensation, like health insurance and retirement plans. Hence, unemployment of labor resources has direct and immediate impacts on the lives of everyone who depends on the productivity of the unemployed worker.

To be unemployed according to the measures used in most industrialized countries, a worker has to be actively seeking work. The unemployment rate measures the percentage of the labor force, which includes those employed plus the unemployed who are actively seeking work but cannot find it. Sometimes the labor force participation rate is a better measure of how intensively productive resources are used in an economy; this measures the percentage of a total population (civilians, noninstitutionalized, over the age of 16) that is either employed or unemployed but actively seeking work. These measures can be applied to subpopulations so that economists can also track unemployment and labor force participation rates by gender, location, racial and ethnic origin, age cohort, and so forth. Given the significance of employment to both individual workers and to national productivity, it is important to watch for trends and understand patterns that might be occurring and their impacts on public policy.

Note that not every member of a population is included in the labor force, and not every person who is pursuing productive activities in an economy is included in the labor force. For example, discouraged workers, defined as people who are not working and have stopped seeking work, are not part of the labor force. People who volunteer in a variety of unpaid capacities, who are working in unpaid internships, or who are engaged in unpaid household or family production, are not included as part of the labor force. Employment data provide a proxy for the extent to which an economy is using its labor resources but come nowhere near truly measuring the output of labor resources in an economy over time.

Unemployment takes a variety of forms and can be considered more or less significant depending on its type. Frictional unemployment tends to be short in duration. When workers lose jobs, it naturally takes some time and effort to seek out and select new job opportunities. Contrast this with structural unemployment, which tends to be long term and occurs because a worker's skills no longer fit the mix of jobs available in the economy in which he or she lives or because a worker lacks the skills needed to find a job that can use them. Seasonal unemployment occurs when the demand for workers of a particular type just does not exist at particular times of the year (there are no blueberries to pick in Maine in January), while cyclical unemployment occurs because demand for labor of many types decreases when the level of economic activity declines (fewer boat salespersons are needed during a recession.)

Figure 14.3 describes the path of unemployment in the United States for the 1999 to 2009 period. It is clear that recessionary pressures in late 2007 and 2008 had a tremendous impact on labor markets and on the number of jobs available to job seekers. Unemployment is referred to as a *lagging indicator* of the level of economic activity, which in this case means that recessionary pressures on other aspects of the economy, like demand for final goods and services or prices of other significant inputs like oil, were impacted months before firms began to lay off workers or decrease hiring. Toward the end of the recession, although other aspects of an economy might be showing marked signs of improvement, the unemployment rates might still be rising. Thus, it is very important that economists and policy makers use movements in unemployment rates with

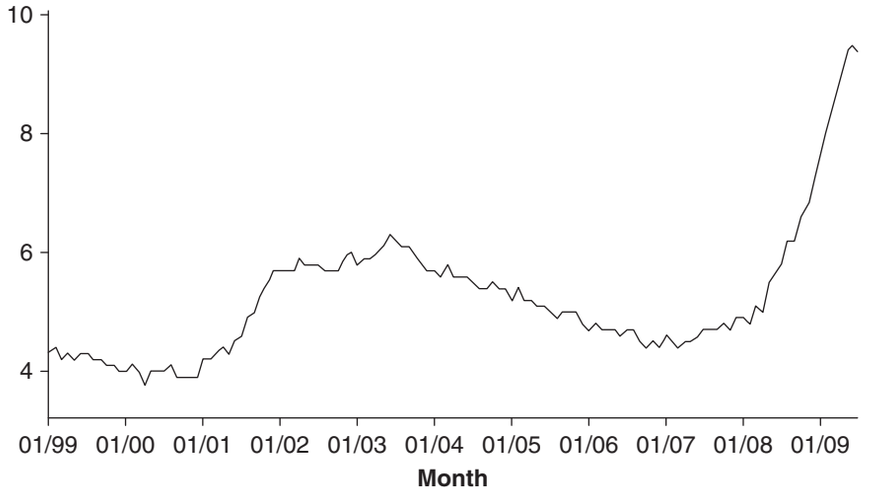


**Figure 14.3** U.S. Unemployment Rate (Seasonalized) 16 Years and Over

SOURCE: U.S. Department of Labor, Bureau of Labor Statistics (n.d.).

extreme care when making predictions about the health of the overall economy.

Table 14.1 describes unemployment rates for a variety of Organisation for Economic Co-operation and Development (OECD) countries. It is clear that industrialized countries have different experiences in their own labor markets. This is caused by a variety of factors, some political and cultural in nature and others having to do with the product mix that the country produces or with historical factors that influence production and consumption. For example, Germany's relatively higher unemployment rates may be due to its recent incorporation of East Germany as part of its economic base or to its generous social safety net that allows the unemployed to receive a greater package of benefits. It could be due to the changes in migration regulations and expectations that have come along with expanding membership in the European Union. It could also be due to the strong manufacturing tradition in Germany that may be increasingly moving to lower-wage countries. All of these sorts of factors must be weighed when comparing unemployment rates across time and across economies.

Though unemployment is always painful for the individual experiencing it, structural unemployment is the type that causes most distress for economists. It leads to an extended lack of income for individuals and their families and can lead to serious psychological problems, including depression, that can lead to other social ills, including domestic violence and substance abuse. The severity of these problems often depends on the social safety provided by the government for the unemployed, which differs widely across countries around the world. Though most industrialized countries have some level of support for unemployed workers, the level of support, the nature of the support (pure monetary support vs. access to job training and/or education), as well as the stigma attached to accessing this support, affect the willingness and ability of workers to remain unemployed.

## Policy Implications

There are many, many public policy implications of labor market decisions and outcomes, some affecting the demand side of the marketplace and some affecting the supply side. Labor markets play fundamental roles in an economy, providing inputs for production, giving people a sense of purpose and well-being, and providing income and economic resources for household consumption. Unemployed workers tend to be angry voters, so most governments around the world embrace full employment, variously defined, as an important component of political and economic stability.

### Employment Policy

Should the goal of an economy be to eliminate all unemployment? Should governments and policy makers establish programs that lead to an unemployment rate equal to zero? Two questions arise here: (1) Is it possible to have no unemployment, and (2) is it desirable to have no unemployment?

**Table 14.1** Comparative Rates of Unemployment Across Industrialized Countries

| | *US* | *Canada* | *Australia* | *Japan* | *France* | *Germany* | *Italy* | *Sweden* | *UK* |
|---|---|---|---|---|---|---|---|---|---|
| 2005 | 5.1 | 6.0 | 5.1 | 4.5 | 9.6 | 11.2 | 7.8 | 7.7 | 4.9 |
| 2006 | 4.6 | 5.5 | 4.8 | 4.2 | 9.5 | 10.4 | 6.9 | 7.0 | 5.5 |
| 2007 | 4.6 | 5.3 | 4.4 | 3.9 | 8.6 | 8.7 | 6.2 | 6.1 | 5.4 |

SOURCE: U.S. Department of Labor, Bureau of Labor Statistics (2009).

Economists sometimes use the term *natural rate of unemployment,* defined as the level of unemployment that will exist in an economy at full employment. This seeming oxymoron is actually a way of describing an economy that has reached a rate of productivity that can be maintained without placing inflationary pressure on resource prices or undue stress on productive resources of all sorts. The natural rate of unemployment in the U.S. economy seemed to be around 5% through the 1990s, but then as the combination of low interest rates and war in Iraq stimulated production, it seemed as though the natural rate fell to around 4.5%. In the United States, because government policy makers and the Federal Reserve Board can both impact the level of overall economic activity, the goal is to achieve a stable level of output and employment in the long run that guides decision making and policy action.

To steer the economy toward full employment, the government can use fiscal policies that affect aggregate demand. In some cases, this means changing the level of spending on some combination of entitlement programs and discretionary spending projects. For example, in 2008, when the U.S. economy seemed headed for a deep recession, the government increased the duration of unemployment benefits, shifted resources into "shovel ready" spending projects like roads and bridges, and promoted the Cash for Clunkers program to increase household spending on new automobiles.

Other government policies directly subsidize job training for workers whose skills do not match current job offerings. This might involve grants to subsidize college students (Pell grants, for example) or more targeted initiatives designed to train workers for occupations that are expected to expand in the near future. For example, the Employment and Training Administration (ETA) administers federal government job training and worker dislocation programs, federal grants to states for public employment service programs, and unemployment insurance benefits. These services are primarily provided through state and local workforce development systems.

In response to the significant challenges presented to American workers by the recession, the American Recovery and Reinvestment Act of 2009 (Recovery Act) was signed into law by President Obama on February 17, 2009. As part of this plan, the ETA will be a key resource for the administration's "Green Jobs" initiative. As described on the ETA Web site (www. doleta.gov),

> The Green Jobs Act would support on-the-ground apprenticeship and job training programs to meet growing demand for green construction professionals skilled in energy efficiency and renewable energy installations. The Act envisions sound and practical energy investments for 3 million new jobs by helping companies retool and retrain workers to produce clean energy and energy efficient components or end products that will result in residential and commercial energy savings, industry revenue, and new green jobs throughout the country.

This type of public policy, directed at workforce development and training, is designed to move workers into productive sectors of the American economy, making important human capital investments that lead to viable employment as well as to a more stable economy.

## Employment Taxes

Income and other employment taxes play an important role in providing incentives for workers to participate in labor markets. One of the primary methods of taxation used by governments at many levels is to directly tax income. Typically, a percentage of each dollar earned is paid as tax, and in many cases the percentage increases as total income increases, making income taxes progressive in structure. This means that workers pay increasing percentages of marginal income as tax as income increases.

Income taxes serve a variety of purposes. First, they provide money to finance government expenditures. Local, state, and federal levels of government all need revenues in order to provide services; income taxes provide that revenue in many cases. Second, income taxes can affect the behavior of workers. If income taxes are increased, workers might increase their participation in labor markets in order to make up for household income lost to tax payments. However, workers might decrease their participation in labor markets in order to take advantage of the now lower opportunity costs of spending time out of the labor markets. That is, because effective wage rates fall as income tax percentages increase, it is less expensive to spend time out of the labor market.

If income taxes are decreased, workers might increase their participation in labor markets because the opportunity cost of spending time out of the labor market has increased. Higher effective wage rates mean that it is more expensive to spend time out of the labor market. However, workers might decrease their participation in labor markets because they can earn the same level of income now with less time spent on the job.

These opposing responses to changes in income tax rates make it very difficult to determine the right mix of policy to achieve government objectives. If the goal is to encourage workers to provide more work hours, should taxes be increased or decreased? If the goal is to encourage workers to provide fewer work hours, should taxes be increased or decreased? Policy has to be very carefully determined, based on the average levels of income earned in a particular market or the historical response of workers to tax changes in the past. For example, typically low-wage workers respond to increases in tax rates by working more hours. If the goal of policy is to encourage workers with jobs to provide a greater number of hours, it is smart to increase income tax rates. On the other hand, if the goal is to encourage labor force participation among discouraged workers, the appropriate policy is to increase the effective wage by lowering tax rates. This means that the opportunity

costs of remaining unemployed have increased and it is now more expensive to stay out of the labor force.

### Minimum Wages

In some instances, government policy makers intervene in markets for a variety of goods and services by setting legal minimum prices. These price floors, set to protect sellers of a good or service, provide sellers with higher levels of income than they would receive if the market equilibrium were allowed to allocate resources. Price floors are used in a variety of markets for goods and services in the United States, particularly in markets for agricultural products like cheese, milk, and sugar.

Minimum wages are used in labor markets for the same reasons. Legal minimum wages provide low-wage workers with protection from wages that may be low because of large supplies of workers who enter these markets. Particularly in urban areas or areas near borders, where large numbers of immigrant workers tend to settle, minimum wages help to alleviate poverty among the nation's most vulnerable populations.

In the United States, the federal government established mandatory minimum wages through the Fair Labor Standards Act in 1938. In the midst of the Great Depression, this policy was designed to provide minimal levels of income for workers who were trying to make their way back into labor markets after extended spells of unemployment. In 2009, the federal minimum wage was increased from $6.55 per hour to $7.25 per hour. Some states, particularly those with very high costs of living, like Connecticut, set their minimum wages higher than what is federally required. Though federal lawmakers have increased the wage several times in recent years, the federal minimum is not indexed to inflation or to increases in labor productivity, and so workers have no guarantee that they will maintain purchasing power over time or that their compensation will rise as their own productivity increases.

Increases in the minimum wage have become quite controversial in most of the countries or markets in which they are imposed. Some argue that establishing wage floors causes higher levels of unemployment in affected markets. For example, some people argue that in the market for people who wash dishes in New York City, if firms are required to pay higher wages, they will move up their demand curves and hire fewer worker hours. Others argue that though this might be true, the demand for dishwashers in New York City is highly inelastic; this means that when wages rise, quantity demanded falls, but by a relatively small amount. So the question becomes an empirical one: When wages rise in low-wage labor markets, how significant is the decrease in quantity of labor demanded? If this decrease is small, then the income gains to workers who remain employed create greater benefits, even though some workers are forced out of the labor market. Many studies have been conducted to measure this impact, particularly on teenage workers, who are the most common recipients of the minimum wage. Though results are mixed, most conclude that the negative impact of increases in the minimum wage on employment in affected markets is small or nonexistent.

A more extreme form of the minimum wage that is gaining momentum is the *living wage.* This is defined as a minimum amount of money required by a worker to maintain his or her own living within a particular market. The Universal Living Wage Campaign is based on the premise that anyone working 40 hours per week should be able to afford basic housing in the market in which that labor is exchanged; this obviously requires higher hourly wage rates far in excess of the federal minimum wage. For example, the living wage in 2002 was $10.86 in New Haven, Connecticut, and $10.25 in Boston, Massachusetts. Even at these higher levels, the price elasticity of demand for labor in low-wage markets is quite inelastic, indicating that increasing wages to even this level will not result in significant increases in unemployment.

## Future Directions

Labor markets are complex and dynamic, and so the possibilities for future directions are endless. The Bureau of Labor Statistics projects significant changes in the labor force in the coming years, as follows. Fewer younger workers will be available to enter the labor force, and most of the growth in the supply of labor will come from new immigrants to the American economy. On the demand side, manufacturing jobs will continue to move overseas, while job growth in the green economy and service sector will continue to increase (Toossi, 2007). Combined with these demographic and sector shifts, several other factors will help to determine the nature of labor markets in the coming decade.

### Technology

As technology provides greater opportunities to enhance worker productivity, it also provides applications that replace worker effort with machines, computers, and robots. This tension between labor and capital as substitutes in production (capital equipment and technology replacing humans in production processes) and as complements in production (capital equipment and technology increasing the productivity of humans in production processes) will not be resolved easily and will need to be considered on a case by case basis across the economy as new technology changes the nature of the work that people do.

### Immigration Policy

As noted above, immigrants are almost certain to account for a significant portion of the growth in the labor force in the next decade. Immigrants enter the United States seeking higher income and living standards than they have experienced in their home countries. Given the

inequality in the distribution of income and wealth around the globe, as long as immigrants can gain access to the American marketplace, the investment in migration is more than worth the costs in terms of personal and family dislocation and downward job mobility.

The key to predicting the impact of this immigration is the degree to which U.S. citizens are able to provide welcoming communities for workers from other countries. Though a variety of laws and regulations limit the ability of firms and communities to discriminate against workers they do not want to accept, equal treatment and opportunity is not the norm in many regions of the country. This means that social networks among immigrant populations become more and more important, and the ability of immigrants to gain access to skills, attitudes, and workplace norms will be crucial if labor productivity is to continue growing as new migrants are absorbed into American life.

The burden of accommodating large immigrant populations can be quite overwhelming to a community. Particularly if border communities like Miami, Los Angeles, and Houston are considered, increases in immigration (both legal and illegal) strain public services like hospitals and schools. Even when people in a community want to welcome productive workers into their midst, they may find it difficult to provide for them in ways that are equitable.

### Labor Unions

The union movement in the United States has declined steadily during the past five decades, with membership down to around 12% of the labor force from around 30% in the late 1950s. There are many reasons for this decline, some economic and some political. The primary sectors of the economy that led the union movement have declined in importance in recent years, led by autoworkers, mine workers, and garment workers. All of these industries have seen increasing competition from international producers and have subsequently been unable to compete with firms in the global economy that have gained a variety of efficiencies in production and employment.

Further, changes in the political climate have made it increasingly more difficult for unions to organize workers. On August 3, 1981, more than 12,000 members of the Professional Air Traffic Controllers Organization (PATCO) went on strike and were subsequently fired by President Ronald Reagan, who determined that the strike was illegal. This decision, in one instant, shifted the balance of power between firms and unions significantly in the direction of employers. From that time, policy and decisions by the National Labor Relations Board (NLRB) began to work in favor of the employer and more often against the ability of workers to form collective bargaining arrangements with their employers. Note that the labor movement in other countries, particularly in Europe, remains strong, with around half of all workers unionized in Great Britain, Germany, Austria, and Italy.

Finally, as the manufacturing sector of the economy has shrunk, the service sector has grown tremendously, representing nearly 80% of business activity in the United States today. Service employees have historically been difficult to organize because they are often female dominated (women are more difficult to organize) and often dispersed across a wide variety of work settings (small offices, working from remote locations, etc.). Unless employees in service industries find ways to unionize more effectively, the union movement will continue to lose relevance in the American economy.

## Conclusion

In a perfect world, labor markets allocate human effort in production toward its most productive use. This chapter has endeavored to explain that allocation process by exploring the behavior of buyers and sellers in markets for human resources and then introduced the role of government policy makers in altering these market-determined outcomes.

The key to understanding the nuances of labor market behavior is in remembering that work is a fundamental source of human dignity. Though economists often focus on labor as a productive resource, which it of course is, there are aspects of the relationship between employer and employee that are clearly emotional, value laden, and culturally determined. This leads to a level of complexity in resource allocation that we do not see with other productive inputs like computers, robotics, or acres of land. However, it is this complexity that provides us as economists with rich avenues for intellectual investigation and policy analysis.

## References and Further Readings

American Association of University Professors. (2006, March-April). Annual report of the status of the profession, 2005–2006. *Academe.* Available at http://www.aaup.org/AAUP/pubsres/academe/2006/MA/sal

Becker, G. S. (1965). A theory of the allocation of time. *Economic Journal, 75*(299), 493–517.

Becker, G. S. (1971). *The economics of discrimination* (2nd ed.). Chicago: University of Chicago Press.

Becker, G. S. (1981). *A treatise on the family.* Cambridge, MA: Harvard University Press.

Bergman, B. (1986). *The economic emergence of women.* New York: Basic Books.

Blau, F. (1984). Discrimination against women: Theory and evidence. In W. A. Darity, Jr. (Ed.), *Labor economics: Modern views* (pp. 53–89). Boston: Kluwer-Nijhoff.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *American Economic Review, 77,* 531–553.

Freeman, R. B., & Medoff, J. L. (1984). *What do unions do?* New York: Basic Books.

Goldin, C., & Katz, L. F. (1998). The origins of technology-skill complementarity. *Quarterly Journal of Economics, 113,* 693–732.

Heckman, J. J. (1974). Life cycle consumption and labor supply: An explanation of the relationship between income and consumption over the life cycle. *American Economic Review, 64,* 188–194.

International Labour Office. (2003). *Yearbook of labour statistics.* Geneva, Switzerland: Author.

Katz, L. F., & Murphy, K. M. (1999). Changes in the wage structure and earnings inequality. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (pp. 1463–1555). Amsterdam: Elsevier.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review, 90,* 1346–1361.

Mincer, J. (1962). Labor force participation of married women. In H. G. Lewis (Ed.), *Aspects of labor economics* (Universities National Bureau of Economic Research Conference Studies No. 14, pp. 63–97). Princeton, NJ: Princeton University Press.

Mincer, J. (1976). Unemployment effects of minimum wages. *Journal of Political Economy, 84,* 87–104.

Nelson, J. (1993). *Beyond economic man: Feminist theory and economics.* Chicago: University of Chicago Press.

Rosen, S. (1986). The theory of equalizing differences. In O. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics* (pp. 641–692). Amsterdam: Elsevier.

Spence, A. M. (1973). Job market signaling. *Quarterly Journal of Economics, 87,* 355–374.

Toossi, M. (2007). Labor force projections to 2016: More workers in their golden years. In *Monthly Labor Review.* Available at http://www.bls.gov/opub/mlr/2007/11/art3full.pdf

U.S. Department of Commerce, Census Bureau. (1975). *Historical statistics of the United States, colonial times to 1970* (Bicentennial Ed., Vol. 1, pp. 131–132). Washington, DC: Government Printing Office.

U.S. Department of Labor, Bureau of Labor Statistics. (n.d.). *Labor force statistics from the current population survey.* Available at http://data.bls.gov/PDQ/servlet/SurveyOutputServlet?request_action=wh&graph_name=LN_cpsbref3

U.S. Department of Labor, Bureau of Labor Statistics. (2007). *National compensation survey: Occupational wages in the United States, 2006* (table 2). Available at http://www.bls.gov/ncs/ncswage2006.htm

U.S. Department of Labor, Bureau of Labor Statistics. (2008–2009a). Physicians and surgeons. In *Occupational outlook handbook.* Available at http://www.bls.gov/oco/pdf/ocos074.pdf

U.S. Department of Labor, Bureau of Labor Statistics. (2008–2009b). Registered nurses. In *Occupational outlook handbook.* Available at http://www.bls.gov/oco/pdf/ocos083.pdf

U.S. Department of Labor, Bureau of Labor Statistics. (2009). *International comparisons of annual labor force statistics, adjusted to U.S. concepts, 10 countries, 1970–2008* (table 1-2). Available at http://www.bls.gov/fls/flscomparelf/unemployment.htm#table1_2

Webster, T. (1999). Work related injuries, illnesses, and fatalities in manufacturing and construction. In *Compensation and working conditions.* Available at http://www.bls.gov/opub/cwc/archive/fall1999brief3.pdf

Viscusi, W. K. (1993). The value of risks to life and health. *Journal of Economic Literature, 31,* 1912–1946.

# 15

# Wage Determination

Kenneth A. Couch
*University of Connecticut*

Nicholas A. Jolly
*Central Michigan University*

Wage determination refers to the market process that establishes the amount a firm pays a worker for a unit of time. A market exists when there is demand for and supply of a product. In the case of the labor market, firms demand individuals' time as an input for production, and workers supply it. The nature of the labor market, including the number of firms and special qualities of workers, influences the wage determination process. Other social institutions such as the government and interactions between individuals, including those characterized by racism and sexism, also influence payments made to workers. This chapter discusses the basic process of wage determination along with the influences of market structure, government regulation, and other social interactions on it.

One reason wage determination receives broad attention among economists is that payments to labor constitute roughly 70% of the gross domestic product of every industrialized country. Most societies would like to put government policies in place to positively influence wage growth because, in aggregate, what people can earn relates directly to the material well-being of a country; however, this requires an understanding of what factors influence the level and growth of wages over time. Perhaps the most important determinant of wage rates is the level of education a person receives; however, theory indicates that general education will have a different impact on labor markets than specific training related to one industry or occupation. Governments attempt to directly affect the mix of general and specific skills among members of the labor force. Thus, studies of wage determination relate directly to choices of broad social policies aimed at influencing individual and social welfare.

Another reason this topic receives substantial attention is that many societies have concluded that payments in the labor market should be determined by individual productivity and not by such factors as gender or race. Laws prohibiting these discriminatory practices reflect this social judgment. Studies of differences in wage rates for people who have the same qualifications but are not of the same gender or race seek to gauge the extent of discrimination in the labor market.

In discussing wage determination, it is important to understand what the term *wage* means. A wage rate is a payment from a firm to a worker for an increment of time. Most commonly, a wage rate is a payment for an hour of effort. However, the term, at times, can capture weekly, monthly, or annual pay. Restricting wages to reflect money paid to a worker for a defined period of effort isolates the payment from its interaction with hours of work in determining earnings, which is a separate concept. Examining the simultaneous determination of the hours individuals choose to work given the wage rate they receive is a standard topic within labor economics; however, it is beyond the purposes of this chapter.

This chapter first discusses how the structure of the labor market determines wages. This section emphasizes that the number of firms demanding labor significantly alters the wages workers receive in the market. It also discusses the role of labor unions relative to large employers.

The chapter then proceeds to discuss how investments in human capital in the form of general education and on-the-job training alter the wage a worker receives. The discussion then moves on to consider different theories of how wages change over the typical person's lifetime. The next major topic is discrimination's affect on wage determination. Finally, the chapter discusses the role of government in setting boundaries on market outcomes, followed by a brief discussion of policy implications and directions for future research.

## Market Structure

This section provides an explanation of how the structure of the labor market affects wages. The discussion begins with a review of how wages are set in a perfectly competitive labor market. One of the most important characteristics of competitive markets is that there are many firms competing for workers' services. When this condition is not present, employers do not face competition for employees and have more freedom to set wages. At the extreme, there is only one firm seeking to hire workers. Economists refer to this type of market as *monopsonistic.* To emphasize the importance of market structure, this section explores the differences between monopsony and competitive markets. The section concludes by discussing how labor unions can counteract the market impacts of large employers as well as other effects attributed to them.

### Perfect Competition

Supply and demand characterize every economic market. Market supply is a summary of the total quantity of a good available at different prices. Market demand is a summary of the willingness to purchase at different prices.

In the case of the labor market, individuals supply, and firms demand, time. Individuals expect payment for their time, and they must decide, given their tastes and other sources of income available to them, how many hours they would be willing to provide at different wage rates. Summed across all workers, this linkage between the differential willingness of workers to provide time to the market as wage rates vary determines aggregate supply in the labor market, the total amount of labor available at different wages. Generally, workers are willing to provide more hours of their time to employers when wage rates are higher. Therefore, aggregate supply increases with the wage rate.

Demand for labor arises from the usefulness of workers to individual firms. Firms require labor to produce goods, and the dollar amount of total production associated with each additional worker is his or her value to the firm. Firms pay for additional labor as long as the dollar value of the additional productivity of another worker exceeds the wage paid for that person's time. As long as the dollar value of the productivity of more labor exceeds the required payment, the firm's profits expand, thus giving it an incentive to hire more. For a given type of labor, summing the demand across firms determines at different wage rates how much labor is used in total. If the wage rate is high, firms hire fewer workers and if low, they use more labor. Therefore, the demand for labor expands as wage rates fall.

Equilibrium in a perfectly competitive market occurs when the quantities of aggregate labor supply and demand are equal. Because the wage rate determines the quantity of labor supplied and demanded, the wage that prevails in a competitive market is such that the number of individuals willing to work equals the number of workers firms wish to hire. Thus, the interactions of aggregate demand and supply determine the market wage rate. This implies that firms pay the same wage rate to all workers. When economists speak of wage determination, at the most basic level, this is the rate of pay for a unit of labor determined by the interaction of market demand and supply.

For this theory to be operational, many assumptions must hold. First, the basic theory applies to one uniform type (homogenous) of labor. The theoretical model also assumes that there are many workers. Therefore, firms can hire as much labor as they would like at the prevailing wage. Theory also assumes that there are many employers competing for the time of the workers and that each firm is small relative to the market. These assumptions imply that neither a single worker nor a single firm has influence to negotiate for higher or lower wages. This is important because it means that if one employer tries to underpay workers relative to the going wage rate, workers can turn to other firms for employment. Free mobility of workers across many employers helps maintain the wage rate in the perfectly competitive model. The equilibrium wage in a perfectly competitive market is known as the *competitive wage rate.*

When any one of these assumptions fails, alternative market structures arise. Because the assumption that fails is associated with a competitive market, it is a case of *market failure* because it represents a departure from perfect competition. To illustrate this point, the next section considers the case of a market with one employer, monopsony.

### Monopsony Markets

The difference between a monopsonistic labor market and a perfectly competitive one is that under monopsony there is only one firm, the monopsonist. One assumption of a competitive market is that all firms are relatively small so that no single employer has influence over the competitive wage. However, if there is only one employer, then hiring decisions of the firm determine the wage paid to workers.

In this market, one firm determines the amount of labor used. As the firm expands its use of labor, it will have to

pay a higher wage rate in order to make more workers willing to supply their time. Assuming the firm pays all equivalent workers the same pay, when it hires one new person, it must raise the pay rate of all current employees by an amount equal to the difference between the wage they were formerly paid and the new, higher one. Correspondingly, the firm views the additional cost of each individual worker hired as being far higher than what it might expect to pay one more person. This consideration leads the firm to hire fewer workers than would be employed in a competitive market. Because the firm uses less labor than would be employed in a competitive market and it can attract smaller amounts of labor at a lower wage rate, workers receive less pay than they would if many employers were active in the market. The result of having one firm hire all workers is lower pay and less employment than would be observed in a competitive labor market.

### Labor Unions

Economists initially developed the monopsony model to explain one-company towns that characterized industries like mining (Pencavel, 1991). A company might locate an iron ore deposit and purchase the rights to it, and as it organized the extraction process, it would be the only firm to hire miners at that location. Individuals viewed these firms as paying workers too little and working them too hard, which is consistent with the general outcomes of the theoretical model. In contemporary markets, economists use monopsony to describe hiring for hospitals in towns where there are no competitors, as well as professional sports where a single league employs all players.

Labor unions have always formed to enable workers to bargain more effectively with very large employers. Unions such as the United Mine Workers and the National Football League Players Association formed to raise more effectively worker concerns with employers than a single individual could. To be most effective, unions must organize the relevant pool of labor for an employer so that they can credibly threaten to disrupt business if the employer does not address workers' concerns.

By organizing workers for an industry, unions have the ability to determine the amount of labor supplied to a firm at a given wage rate. In the monopsony model, a firm employs fewer workers because the cost of hiring an additional worker is larger than in a competitive market. Unions have the ability to set a uniform wage rate equal to what it would be in a competitive market through bargaining, and this can potentially raise both wage rates and employment back to the competitive level.

These positive views of unions are tempered by the observation that as workers vote on contracts, older workers are protected under union rules by seniority from being fired and are thus more likely to support demands and vote for pay packages that set wages above the level that might exist in a competitive market. To the extent that this is true, because the demand for labor declines as wages rise, this will reduce employment to below the level expected in a competitive market.

Unions receive a great deal of scrutiny for their impact on reducing the competitiveness of firms. If unions raise wages above levels observed in nonunion enterprises, then they arguably place the firm at a price disadvantage unless the higher cost is justified by increased productivity. Only 15% of all full-time workers over the age of 25 belonged to a union in 2000. By 2007, this percentage declined to 13.3% among all workers and was less than 10% among private sector employees. Because of this limited coverage, unions have less of a role in wage determination in contemporary labor markets than in times past. See William Dickens and Jonathan Leonard (1985) for an analysis of the determinants of the decline in union membership.

## Human Capital Acquisition

The previous section emphasized how market structure and worker responses to it affect wage determination. Rather than acting collectively, workers may make individual efforts to raise their rate of pay by investing in themselves to increase their productivity. Early work by Gary Becker (1962) and Jacob Mincer (1974) first conceptualized the idea that individuals may invest in themselves, hoping to recoup that expense through increases in their rate of pay. Thus, individuals may choose to invest in their human capital much as firms do for physical machinery.

Although most people think of investments in human capital as originating from the individual, firms have an incentive to invest in workers when this helps tie workers to them. Thus, the basic theory distinguishes between general skills that are transferable across firms and specific skills that raise productivity only for a specific employer. When skills are transferable to other firms, employers will not finance their acquisition. If skills are specific to a single employer, then firms should finance their acquisition.

The discussion here proceeds assuming that formal education leads to the formation of general skills that are transportable across employers. The reader should bear in mind that firms do have an incentive in some circumstances to invest in formal education.

### Formal Education

Education increases worker productivity. Because increased productivity brings value to the firm, education should increase the pay an individual receives. In deciding whether to pursue a course of education, such as a bachelor's degree, an individual must examine not only the short-term earnings impact, but also the long-term effect of education on earnings. If the person is going to pursue a degree at a university, then most of the costs will occur

in the early years and increased earnings will come later. So the individual should consider what the total stream of earnings from the additional education would be over his or her lifetime in current dollars. The individual must also consider what earnings would be in current dollars over a lifetime without the investment. By making a direct comparison of the earnings paths with and without the additional investment in terms of current dollars, individuals can calculate the rate of return that would occur if they spend money on raising their education level. Because individuals can invest their money elsewhere, a rational person should compare the rate of return from pursuing education to possible returns from alternative investments. Individuals will pursue education as long as the rate of return on that investment is at least as large as from other opportunities.

Becker (1962) also outlined two major factors thought to impact individual investments in education. The first is ability. If people differ in their ability to process information, then when exposed to the same material, they will carry away different levels of comprehension and effectiveness in transforming the information into productive workplace skills. Individuals should consider their ability levels as they make decisions about how much education to pursue. Students receive a considerable amount of information on their academic ability. Grades and standardized tests are routine. Theory suggests that students who have received feedback that their academic ability is relatively low would invest less in education because it would not raise earnings as much as for those with greater ability.

The other major factor that influences the amount of education pursued is the availability of financing. Many students take loans to pursue further years of education, while others receive grants, scholarships, or financial assistance from their family. If a person receives money to help pay for education through scholarships, grants, and no/low-interest loans, then the cost of education decreases and the rate of return increases. Reducing the cost of education in this manner encourages higher amounts of educational investments.

Social observers are concerned that these two factors are positively correlated. Students who have stronger backgrounds and higher ability may also be more likely to come from families with more resources. Weaker students may come from families that have fewer resources. Because stronger students would pursue more education in the absence of free financing and weaker students less, a positive correlation of ability and financing reinforces those patterns.

In terms of influencing wage determination, those with more formal education generally receive higher wages because they possess more skills. Employers look at courses of education as one factor that determines whether a worker carries useful skills. Because workers who are more skilled receive higher pay, many countries emphasize improvements in their national education systems as a method of raising average rates of pay and material well-being.

## On-the-Job Training

In some cases, firms may have an incentive to invest in general skills of workers through on-the-job training. However, most employers use this method to teach skills that are useful only in their own firm. If firms did invest in general education for workers, the employees would receive a wage rate low enough to pay fully for the training. Here, the discussion focuses on training in specific skills. The seminal work of Becker (1962) and Mincer (1974) also explains the role of on-the-job training.

Specific training increases workers' productivity with their current employer after it is completed. From the perspective of the firm, if the worker quits after receiving training, then the firm absorbs the lost cost of instruction. From the employees' perspective, if they take reduced pay during the training period, and the firm fires them, then the workers have made a human capital investment that is worthless to other potential employers. Because both parties have something to lose, they need a positive incentive to participate in on-the-job training.

The firm and the worker share the risks of on-the-job training by splitting the costs and the benefits from it. With specific training, workers earn lower wages than they might receive elsewhere during the training period. However, as long as the wages are reduced by a fraction of the training costs, the worker shares only part of the cost of investment. So the firm and worker bear part of the cost. After training is completed, workers' pay remains below the actual value of the increased output associated with their new skills, but they will receive a wage that is larger than what they would earn at a different employer both at that moment and over their career. Thus, both the worker and firm have a positive incentive to participate in specific training.

## General and Specific Training and Displaced Workers

To examine empirically how general and specific human capital affects wage determination, economists examine the experiences of displaced workers. Job displacement refers to the loss of employment due to plant closure or large-scale layoff. While it is possible that employers use layoffs to purge problematic workers from a firm, when layoffs are of a massive scale or when an entire firm closes, economists view this type of event as beyond the control of the worker. Because of this, researchers treat events like plant closure as naturally occurring social experiments. The workers who lose jobs change firms. Thus, the component of wages related to specific firms is lost. By comparing displaced workers' wages to those of individuals who remain continuously

employed, researchers can get an idea of the proportion of wages related to specific ties to individual firms.

Empirically, when workers experience job displacement, earnings drop significantly. In summaries of the literature, Lori Kletzer (1998) and Bruce Fallick (1996) report that earnings are below where they would be if displacement had not occurred for several years afterward; however, earnings usually recover to within 10% to 15% of their prior level in a 5-year period. The magnitude of displaced workers' long-term earnings losses provides an estimate of the proportion of wages related to firm-specific factors.

Because the U.S. education system emphasizes the acquisition of general skills, one might be interested in knowing whether this small proportion of skills arising from attachment to specific employers is unique to the United States or if other countries exhibit this pattern. There is reasonable evidence from France (Abowd, Kramarz, & Margolis, 1999) and Germany (Couch, 2001) that the proportion of pay attributable to specific employers is in the range of 10% to 20% as is the case for the United States (Couch & Placzek, in press).

## Why Are Age-Earnings Profiles Positively Sloped?

The method of discovery in any science, including economics, involves formulating theories meant to be useful in explaining empirical observations. Researchers gauge the validity of competing theories by examining whether they are consistent with empirical reality.

Human capital theory (Becker, 1962; Mincer, 1974) originally sought to explain why wages grow at younger ages, peak at midlife, and decline afterward. The extension of human capital theory to life cycle considerations by Yoram Ben-Porath (1967) predicts that early in life people should specialize in acquiring general skills because forgone earnings are smaller costs of acquiring education at younger ages. As individuals enter the labor market and obtain general skills from work and specific skills from employers, wages will grow over time. Eventually, as workers' energy declines and reinvestments in skills taper off, wages fall. Finally, the worker exits the labor market.

The previous section described the proportionate contribution of general and specific skills to wage growth. However, economists debate about how strong the linkage is between wage growth and worker productivity at any given point in time. There are two major alternative theories of why wages grow over time: efficiency wages and job matching. Efforts to test the theories simultaneously in order to determine which effect is dominant are ongoing in the profession.

### *Efficiency Wages*

In many workplaces, employers face difficulty in knowing whether their employees are putting forth the appropriate level of effort. When employees use paid time for nonwork activities, they are shirking their responsibilities. One way employers respond to this behavior is by using wages to encourage greater effort, or efficiency, from workers (Krueger & Summers, 1988).

Early in an individual's career, a firm might choose to pay wages that are smaller than the actual value of the person's product. This can be an effective device for learning which employees are most committed to the firm. However, if workers do not expect full payment for their efforts, they will move to a higher-paying employer because the total stream of compensation would be larger. To offset this concern, the firm must overpay the workers later in their career. Additionally, these theories require the firm to have the ability either to terminate older workers so that total compensation over their lives does not exceed the value of their output, or to use other mechanisms such as the structure of pension payments to induce retirement of workers at a desired point. The essential distinction between this and human capital theories is that over the majority of the individual's life cycle, wage rates are not tied to the dollar value of worker productivity at a point in time.

One common example of a labor market with efficiency wages is academics. Younger faculty are often more productive than their older counterparts. Whether a new faculty member will receive tenure (a guarantee of employment) is normally determined 6 years after beginning employment. A university may be concerned that once a professor receives tenure that person will reduce his or her level of productivity. For the first 6 years, the university underpays and overworks junior faculty. The university heavily scrutinizes their productivity records. This screening is arguably sufficient to assure that the person being evaluated is devoted enough to research and teaching that he or she will remain productive after receiving tenure. Often, when a professor receives tenure, pay increases. When older professors are not as active in producing research, they receive wages that are large relative to their productivity.

The above example describes how efficiency wages work to increase productivity over time. Firms also use efficiency wages in a static context to increase worker productivity. In that case, efficiency wages are payments by firms structured to provide incentives to workers to be more productive or efficient in their efforts. Examples of efficiency wages include payments aimed at maintaining a healthy workforce, attracting more productive workers, and keeping employees from shirking their responsibilities.

### *Job Matching*

Job matching theory provides an alternative view of wage growth. According to this theory, workers seeking employment receive job offers. As workers and firms

match, they both need to discover whether the match is good. This discovery process takes time. During this process, workers are willing to accept lower wages in the hopes that the match is good. If firms and workers decide that the match is good, then wages increase and the pay is higher because the firm finds that the worker is valuable. Because the worker is happy there, the employment relationship is more durable. Thus, this theory predicts that wages grow over time primarily because of a good initial match with a firm. Empirical evidence from Robert Topel (1991) supports this theory because it finds that neither the length of time a person has spent in the labor market nor the length of time a person has worked for a particular employer is systematically reflected in wages.

## Discrimination

Generally, there are three actors in the market: workers, firms, and customers. Becker (1971) explained how in a complete economy discriminatory behavior by any one of these actors affects wages. He defined discrimination as what occurs when someone in a market is willing to pay a price for prejudice against a person or group. Prejudice in a labor market entails treating someone differently because of nonmarket factors unrelated to productivity, such as gender or race.

If employers discriminate against a group, then they will act as if the wage they must pay for a worker has an extra price embedded in it. This means that if a firm hires someone it dislikes, then that worker must accept lower pay.

Similarly, if customers are prejudiced, then they will pay a higher price for a product rather than do business with someone they dislike. Some businesses counter customer prejudice by hiring minorities for jobs that are not visible to the public. For minority-owned firms to gain business of prejudiced customers, they must accept reduced payment for their services.

Workers themselves can also be prejudiced. There are jobs that some think of as appropriate only for a certain type of worker. For example, many think of caregiving jobs as being most appropriate for women. Social pressure exists for women to accept caregiving roles not only in life but also in their choice of work (Friedan & Quindlan, 1964). When this type of pattern arises, workers can be crowded into specific occupations. This oversupply of labor depresses wages in those markets.

### Measuring Discrimination

Ronald Oaxaca (1973) developed a method to measure the impact of discrimination in labor markets. This method estimates how much of a pay gap between minority workers and a comparison group (usually prime-aged white males) arises because of observed factors. After accounting for observed differences in productivity, the remainder, or residual, is interpreted to be an upper bound on discrimination's effect on wages.

Beginning in the 1970s, wage inequality increased in the United States for all groups of workers. According to research by Gottschalk and Moffitt (1994), approximately half of the increase in inequality is unexplained. Using Oaxaca's method, this increasing inequality would be included in the residual or unexplained category. To correct this, Chinhui Juhn, Kevin Murphy, and Brooks Pierce (1993) extended Oaxaca's method so that general increases in wage inequality common to all workers would not be attributed to discrimination.

### The Black-White Pay Gap

Kenneth Couch and Mary Daly (2004) show that for both recent entrants to the labor market and prime-aged workers, the overall black-white pay gap in the United States has declined to historic lows. For prime-aged workers, the overall pay gap remains at about 25%, with about half of it unexplained. This unexplained component provides an upper bound on discrimination's impact on pay. For those with less than 10 years of experience, the pay gap is around 15%, with about 80% of the gap unexplained. For both prime-aged and younger male workers, discrimination appears to reduce wage rates of blacks by approximately 12 percentage points.

### The Male-Female Pay Gap

In recent estimates of gender discrimination's impact on wages, Louis Christofides, Qi Li, Zhenjuan Liu, and Insik Min (2003) report that among participants in the labor market, women's pay remains at about a 26% deficit relative to men. Nine percentage points of that gap are explained by observable differences between women and men.

## Policy Implications

The government takes two broad philosophical approaches to intervening in markets. The first approach entails identifying a market failure and trying to correct it through regulation or taxation. The second approach arises when there is no specific market failure but the competitive outcome is socially unacceptable and needs alteration even if there are costs, in terms of economic efficiency, of doing so.

It is not always easy to determine which perspective is behind government action because most intervention in economic affairs requires agreement of many people holding different views. Nonetheless, it is clear that the government is active in managing rates of pay, market structures, educational attainment, and discrimination in the labor market.

## Rates of Pay

Since the passage of the Fair Labor Standards Act in 1938, the U.S. Congress has set minimum rates of pay for workers in specific industries and occupations. Proponents of minimum wage statutes argue that some jobs are beneath the dignity of American workers (see the discussion in Burkhauser, Couch, & Glenn, 1996). According to this view, even if raising wages reduces demand for workers and unemployment rises, losing those jobs is appropriate because Americans should not work for such low pay. Others (Card & Krueger, 1997) have argued that imposing minimum wage standards is inconsequential because this legislation has no adverse affects on employment. However, this is a point of contention in the profession (Burkhauser, Couch, & Wittenberg, 2000).

Some argue that there are more effective ways than legislating minimum wages to get money into the hands of the working poor. This is because many who work for the minimum wage are teenagers or secondary earners in the household. For example, for every dollar of cost, direct tax rebates put approximately $0.85 into the hands of a low-income household versus $0.15 out of every dollar of a minimum wage increase (Burkhauser, Couch, & Glenn, 1996). Regardless of whether one agrees that minimum wages reduce employment or truly assist the working poor, it is undeniable that federal and state governments intervene directly in the labor market.

## Market Structure

The primary motivation for regulating industrial structure is to prevent the deleterious effects that arise from having too few firms in a market. The U.S. government assesses whether mergers of large corporations would create an anticompetitive environment within an industry. The authority to assess the appropriateness of firm mergers is established through statutes, including the Sherman Antitrust Act, the Clayton Act, and the Federal Trade Commission Act.

Most researchers who study the impacts of antitrust regulation focus on the impacts on consumer prices or the negative effects for U.S. firms in international markets if the firms' sizes are insufficient to compete effectively. As discussed previously, another positive impact of retaining a large number of smaller firms in an industry is a more competitive market for workers. Economic theory suggests markets with more firms will tend to pay higher wages.

## Educational Policy

Because of its role in developing worker skills, the formal educational system in every country is linked to the labor market. At the secondary level, some courses impart general skills useful to anyone as they move toward becoming a functioning adult in society. Other courses impart general skills while tracking students into college. Students who are not interested in postsecondary education and have specific job interests receive training through vocational courses designed for specific occupations (automotive technician, electrician) while still in high school.

Across countries, the amount of emphasis placed on each of these components varies. Each country makes decisions regarding the educational system's structure, intending to promote wage growth. For example, in the United States, roughly twice as many students as in Germany attend postsecondary education. Few students at the secondary level in the United States participate in education aimed at a specific occupation, while formal occupational apprenticeships are normal in Germany among high school students not tracked toward college (Couch, 1994). The advantage the United States hopes to gain by having more workers trained with higher levels of general skills is greater worker flexibility. Germany hopes to have workers who are more productive because they are more specifically trained for their occupation. Germany hopes that the higher productivity arising from more skills that are specific will justify higher wages and result in more vital, competitive firms. A consensus has not evolved as to which system is better.

Countries also vary in their approaches to reducing the influence of family wealth on educational attainment. One's birth family is a matter of luck, and denying an able student access to education because of lack of financing is economically inefficient because a valuable resource is underused. Most would say that to limit a person's outcomes based on his or her birth family is also unfair. In the United States, the system of universities consists of both private and public institutions, most of which charge tuition. To increase access, government loans and grants are available to students, and the terms and amount of financing are related to family background. In many other countries, the university system is public. Once a student graduates from high school and qualifies for admission to a university, the government pays tuition and other expenses. Whether the education system is public or private, the availability of courses and the extent of public subsidy reflect considerations of economic efficiency, fairness, and long-term economic welfare.

## Discrimination

When someone is treated differently because of color or gender rather than productivity, it is understandable that inefficiency results. If the most able person to complete a job is from a minority group and prejudice relegates that person to some other occupation, then a loss of productivity results. Moreover, differential treatment based on race or gender is patently unjust.

For reasons of justice as well as economic efficiency, the government plays an active role in ensuring that persons

of equal capability receive equal treatment. In the United States, most of the laws that prevent discrimination arose in the 1960s. The Equal Pay Act of 1963 established that men and women working in jobs that require "equal skill, effort, and responsibility" receive equal pay. The Civil Rights Act of 1964 made it illegal to discriminate in employment practices based on "race, color, religion, sex or national origin." Executive Order 11246, signed by President Lyndon Johnson in 1965, prohibited employment discrimination by contractors performing work for the federal government. It also resulted in the establishment of the Office of Federal Contract Compliance, which oversees the affirmative action plans of employers doing business with the federal government.

After these measures were enacted, racial pay inequality declined rapidly in the United States (Couch & Daly, 2004); however, concurrently, racial differences in employment widened (Fairlie & Sundstrom, 1999). It is hard to understand why the proportion of minorities employed would fall and proportion of minorities unemployed would rise in a period when relative pay rose. Recently, this same pattern has arisen again. Racial pay inequality declined in the 1990s, yet observable factors cannot explain the racial difference in unemployment (Couch & Daly, 2004; Couch & Fairlie, 2008). Two principal explanations have emerged for this phenomenon. Some think that discrimination that used to be operational in steering minorities to lower-paid occupations in a firm and denying promotional advances may have moved into the hiring decision (Couch & Fairlie, 2010). Others argue that expansion of government assistance programs has raised unemployment among less-skilled workers (Murray, 1986).

Gender pay differences have been more static over time. Some think that this is due to the passage of laws aimed at reducing racial discrimination at the time when women from the baby boom generation were entering the labor force. There is some evidence that to meet hiring goals for minorities, employers substituted African Americans for women. Additionally, even though average family sizes have declined, women are more likely to have interruptions in their labor market experiences relative to men, and this would reduce their value to employers. These factors help explain, in part, why the observed male-female pay gap for people with comparable education and experience remains persistent.

## Future Directions

While research has lead economists to understand a great deal about the processes by which wage rates are set and the broad set of influences on wage rates, much remains to be learned. For example, economists have yet to determine which theory of wage growth best describes the operation of the labor market. The reason that understanding this is important is that if much of wage growth is determined by being matched into a good job rather than from investing in the expensive process of skill formation, then from a policy perspective, greater effort should be devoted to getting workers into the right firm.

Similarly, while a great deal has been learned about the role of prejudice in the labor market, a clear explanation has not emerged as to why relative pay rates of minorities and whites have converged but rates of unemployment have diverged. Some research points to factors that are likely to be part of the explanation, but no convincing, comprehensive explanation has emerged.

## Conclusion

The rate of pay for workers is one of the basic determinants of individual well-being. The combination of wages and hours worked ultimately determines workers' earnings. Admittedly, most individuals do live in a familial context that also affects their economic circumstances. Nonetheless, the wage rate individuals can earn, along with the time they have available for work, places a fundamental constraint on their consumption possibilities.

Average wages place a similar constraint on societal well-being. Government action can influence many of the factors that affect wage rates. The government has determined that some wage rates are too low to be socially acceptable and has set limits. Similarly, an industry with too few firms deviates in important ways from a competitive market structure, and the government reviews corporate mergers likely to affect adversely the relevant industry. The skills that individuals would like to develop to raise their own productivity are seen as vital to economic progress, and every country manipulates its educational policy to try to assist skill development among its citizens. Having all citizens participate in the economy on an equal basis also promotes economic efficiency, and governments routinely take measures to make opportunities uniform for all citizens.

## References and Suggested Readings

Abowd, J. M., Kramarz, F., & Margolis, D. (1999). High wage workers and high wage firms. *Econometrica, 67*(2), 251–333.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy, 70*(5), 9–49.

Becker, G. S. (1971). *The economics of discrimination.* Chicago: University of Chicago Press.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of Political Economy, 75*(4), 352–365.

Burkhauser, R. V., Couch, K. A., & Glenn, A. J. (1996). Public policies for the working poor: The earned income tax credit versus minimum wage legislation. *Research in Labor Economics, 15,* 65–109.

Burkhauser, R. V., Couch, K. A., & Wittenberg, D. C. (2000). A reassessment of the new economics of the minimum wage literature using monthly data from the CPS. *Journal of Labor Economics, 18*(4), 653–680.

Card, D., & Krueger A. B. (1997). *Myth and measurement: The new economics of the minimum wage.* Princeton, NJ: Princeton University Press.

Christofides, L. N., Li, Q., Liu, Z., & Min, I. (2003). Recent two-stage sample selection procedures with an application to the gender wage gap. *Journal of Business & Economic Statistics, 21*(3), 396–405.

Couch, K. A. (1994). High school vocational education, apprenticeship, and earnings: A comparison of Germany and the United States. *Vierteljahrshefte Zur Wirtschaftsforschung,* Heft 1.

Couch, K. A. (2001). Earnings losses and hours reductions of displaced workers in Germany. *Industrial and Labor Relations Review, 54*(3), 559–572.

Couch, K. A., & Daly, M. (2004, Fall–Winter). The improving relative status of black men. *Journal of Income Distribution,12*(3–4), 56–78.

Couch, K. A., & Fairlie, R. (2010, February). Last hired, first fired? Black-white unemployment and the business cycle. *Demography, 47*(1), 227–247.

Couch, K., & Placzek, D. (in press). Earnings impacts of job displacement revisited. *American Economic Review.*

Dickens, W. T., & Leonard, J. S. (1985). Accounting for the decline in union membership, 1950–1980. *Industrial and Labor Relations Review, 38*(3), 323–334.

Fairlie, R. W., & Sundstrom, W. A. (1999). The emergence, persistence, and recent widening of the racial unemployment gap. *Industrial and Labor Relations Review, 52*(2), 252–270.

Fallick, B. C. (1996). A review of the recent empirical literature on displaced workers. *Industrial and Labor Relations Review, 50*(1), 5–16.

Friedan, B., & Quindlan, A. (1964). *The feminine mystique.* New York: Dell.

Juhn, C., Murphy, K. M., & Pierce, B. (1993, June). Wage inequality and the rise in returns to skill. *Journal of Political Economy, 101.*

Kletzer, L. G. (1998, Winter). Job displacement. *Journal of Economic Perspectives, 12*(1), 115–136.

Krueger, A. B., & Summers, L. H. (1988). Efficiency wages and the inter-industry wage structure. *Econometrica, 56,* 259–293.

Gottschalk, P., & Moffitt, R. (1994). The growth of earnings instability in the U.S. labor market. *Brookings Papers on Economic Activity, 25*(2), 217–272.

Mincer, J. A. (1974). *Schooling, experience, and earnings.* New York: Columbia University Press.

Murray, C. (1986). *Losing ground: American social policy 1950–1980.* New York: Basic Books.

Pencavel, J. (1991). *Labor markets under trade unionism.* Cambridge, MA: Basil Blackwell.

Oaxaca, R. L. (1973, October). Male-female wage differentials in urban labor markets. *International Economic Review. 14*(3), 693–709.]

Topel, R. (1991, February). Specific capital, mobility, and wages: Wages rise with job seniority. *Journal of Political Economy, 99,* 145–176.

# 16

# Role of Labor Unions in Labor Markets

Paul F. Clark and Julie Sadler
*Pennsylvania State University*

Labor unions are organizations formed by employees for the purpose of using their collective strength to improve compensation, benefits, and working conditions through bargaining; to bring fairness to the workplace through the provision of due process mechanisms; and to represent the interests of workers in the political process. Economists have traditionally viewed unions as functioning as labor market monopolies. Because they raise wages above the competitive levels set by the market, economists argue that labor unions create inefficiencies resulting in the loss of jobs and in greater income inequality in the workforce. For this reason, economists view unions as an undesirable interference in the operation of the market (Booth, 1995; Friedman & Friedman, 1980; Simons, 1948). However, some economists argue that in addition to their negative monopoly face, unions have a second, positive collective voice face. They further argue that, on balance, the positive impact of unions outweighs the negative (Bok & Dunlop, 1970; Freeman & Medoff, 1984; Reynolds & Taft, 1956).

This discussion focuses on the role unions play—and the impact they have—in contemporary society and in the labor market. The chapter first examines the historical development of American unions. Next, it discusses the structure and government of modern unions and membership trends. The industrial relations process through which unions advocate for their members is outlined. Finally, the chapter examines the impact of unions and evaluates the two faces of unionism.

## Why Do Unions Exist?

Unions are formed by employees who desire to improve their compensation, benefits, and working conditions and to bring greater fairness and due process to their workplace. Employees recognize that unless they have unusual or unique skills or talents, individuals have very little influence with their employers and very little power to improve the conditions under which they work. However, by banding together, workers are able to exert collective pressure that is more likely to force an employer to make specific improvements in the workplace. This collective power can also be used to obtain improvements through the political and legislative processes.

## What Do Unions Do?

Economists generally assert that when employees are dissatisfied with their jobs, the only rational option available to them is to exit or quit their job and reenter the labor market to seek a better situation. However, some economists recognize that employees have a second alternative. Rather than exit their workplace, they can engage in *voice.* Engaging in voice involves trying to convince an employer to make changes in the workplace that will address employee dissatisfaction.

Employees readily understand that if they engage in voice behavior as individuals, they will probably have little success in convincing their employers to make significant

changes. Even if an employee threatens to quit, an employer can, in most cases, easily replace one worker. Employees also recognize that if they combine their individual voices with those of the other employees in their workplace, they will be significantly more likely to persuade an employer to make changes. Unions are the mechanisms employees form in order to use their collective power.

Over time, employees have come to use the collective power of unions in three different ways. First, they engage in collective bargaining with employers in an effort to establish a legally binding contract that details the compensation, benefits, and working conditions in their workplaces. The goal of such a process is to negotiate terms that are better than those set by the labor market. Unions use the threat of a strike (a work stoppage that puts economic pressure on an employer by halting the production of a product or the provision of a service) to push employers to make improvements in these areas.

Second, unions establish processes and mechanisms through which employees can have a greater say in decisions that affect them. These processes can take many forms, including the establishment of grievance procedures that give employees due process when disciplined and employee involvement programs, such as labor–management committees and quality-improvement plans. All of these processes are established through negotiations between the employer and the union and ultimately give employees an opportunity to voice concerns and participate in decisions about how work will be organized.

Third, the collective power of unions can be used to shape the legal and political environment in which they operate. Unions have long understood that forces beyond the employer and union relationship affect the union's ability to represent its members' interests. To varying degrees, unions have actively worked within the legal and political realms to mold this environment, as well as to obtain workplace improvements through the political and legislative processes. The local, statewide, and national membership of a union represents a significant voting bloc. Unions are able to provide campaign support for local, state, and federal politicians. In addition, unions help register people to vote, provide information to members and the public regarding politicians' stance and record on employment-related issues, raise funds for candidates, and engage in get-out-the-vote initiatives. Using these financial and organizational resources, unions work "to reward labor's [political] friends, and punish its [political] enemies" (Gould, 2004, p. 2). They also actively lobby for legislation that will benefit their members and against legislation they believe will be detrimental to their interests.

## Historical Development of Unions

Most labor historians generally agree that the first real union in the United States was formed by shoemakers (then called *cordwainers*) in Philadelphia in 1792. By 1806, the courts banned the cordwainers, and all other unions, as conspiracies in restraint of trade. When this decision was overturned in 1842, the government found other ways to discourage the formation of unions, including issuing injunctions and the use of police and militia to put down strikes, demonstrations, and other forms of collective action. Meanwhile, most employers used a variety of strategies to fight the unionization of their workforce, including firing union activists and blacklisting them so they could not find other employment, evicting pro-union employees and their families from company-owned housing, and engaging in violence against workers who tried to organize.

These efforts on the part of employers did not entirely prevent workers from forming local, and even national, unions, but they did succeed in keeping unions relatively weak and on the defensive. However, legislation passed in the midst of the Great Depression as part of the New Deal changed the legal status of unions and ensured their existence as a central part of the U.S. economy. The National Labor Relations Act (NLRA), passed in 1935 as part of President Franklin Roosevelt's New Deal, changed public policy toward unions in a dramatic, even radical, way. After more than a century in which government often assisted employers in the suppression of unions, the NLRA now granted most American workers in the private sector a legal right to organize, bargain collectively, and engage in strikes, if they choose to do so. Employers vehemently opposed the legislation, but it was declared constitutional by a 5–4 vote of the Supreme Court. The passage, and confirmation of the constitutionality of the NLRA by the Supreme Court in 1937, meant that the federal government would no longer side with employers against unions; rather, it would actively protect workers' rights to organize, bargain, and strike. The NLRA included a set of rules that would dictate how employers and unions would deal with one another in the future. The violence and chaos that characterized union–employer relations for more than a century was replaced by a systematic process of industrial relations that is still in use today. And while the relationship between unions and management is still adversarial, the inherent conflict between the two parties is worked out under the rule of law, with strikes by unions and lockouts by employers used minimally and as a last resort.

One other major historical development that played a significant role in the formation of the contemporary American labor movement was the rise of public sector unions in the 1960s and 1970s. While the NLRA granted the right to organize, bargain, and strike to millions of American workers, it did not extend those rights to employees working for federal, state, or local government. These public employees did not begin to gain such rights until more than 25 years after the act's passage.

Federal workers were the first government employees to gain the right to organize and bargain collectively. However, the bargaining rights extended to them under Executive Order 10988 in 1962 were (and are) limited. For

example, federal employees were not granted the legal right to strike.

The rights of state and local government employees to form unions and bargain are granted on a state-by-state basis. A majority of states grant public employees the right to organize a union. Most also extend some rights to bargain collectively to those who organize a union (although some states, mostly in the South, do not). But only 11 states allow even some of their state and local government employees the right to strike. And while many states grant public safety employees (i.e., police, firefighters, and prison guards) some bargaining rights, none grant these workers the right to strike.

## Contemporary American Unions

Unions are part of the economic framework of most developed, and many developing, nations. This is particularly the case in Western industrialized countries. However, the unions that have developed and operate in the United States over the last two centuries are somewhat unique compared to their counterparts in other nations. One of the most significant ways that U.S. unions have differed historically from unions in other countries is their acceptance of capitalism and rejection of socialism. Socialism has long been an integral part of the labor movements that developed in the United Kingdom and Ireland, western Europe, and Australia and New Zealand over the past 200 years. Even Canadian unions have had a tradition of supporting socialist principles. This uniqueness is sometimes referred to by historians as *American exceptionalism* and is a function of the American labor movement's acceptance of capitalism (Lipset, 1997).

At various points in America's history, there have been labor organizations that have promoted a socialist ideology. In the late 1800s, the leaders of the Knights of Labor led a quasi-socialist movement that attempted to organize workers on a massive scale. Initially, American workers flocked to the organization, and by 1886, the Knights had grown to 700,000 members. However, the membership quickly became disillusioned by the organization's lack of tangible gains, and by the late 1890s, the Knights of Labor had largely disappeared.

In the early 1900s, another socialist union, the Industrial Workers of the World (IWW, also known as the *Wobblies*) was founded. The Wobblies were even more committed to socialism than the Knights of Labor and much more militant. Their union was based on the principle that "the working class and the employing class have nothing in common" and their stated goals were the "overthrow of capitalism" and the "seizure of the means of production" (Dubofsky, 2000).

While the IWW did attract over 100,000 members by 1917, American workers were fundamentally uncomfortable with their brand of revolutionary, socialist unionism. This, combined with heavy-handed attacks on the organization from federal, state, and local governments, led to the demise of the Wobblies as a labor union of significance (Dubofsky, 2000).

Historians suggest that the collectivism promoted by these two particular groups did not win the hearts and minds of the vast majority of American workers because of the uniquely American culture and worldview that were emerging in the late 1800s and the early 1900s. This culture emphasized individualism, egalitarianism, and a general suspicion of government, all of which were consistent with capitalism and at odds with socialism. If American workers had not embraced these capitalistic principles, the American labor movement could have taken a very different shape (Lipset, 1997).

### Union Membership and Density

As U.S. unions developed during the twentieth century, both the number of union members and union density (the percentage of the labor force that belongs to a union) in the United States fluctuated over time. Figure 16.1 depicts the changes in union density.

From a high of 35.5% in 1945, union density fell over the next six decades to a low of 12.1% in 2007. Among the factors cited for this decline are structural changes in the economy (particularly the shift from an industrial-based to a service-based economy that occurred in the last 20 years), increasingly aggressive efforts by employers to fight unionization, weakening labor laws, and changes in the public's attitudes toward unions. The overall union density rate includes employees in both the private and the public or government sectors. However, the trends in these two sectors have been very different over the last 40 years. Although private and public sector union density rates were both around 25% of the labor force in the mid-1970s, in the decades that followed, public sector union density rose to nearly 40%, while private sector union density rates fell below 10% (see Figure 16.2). The higher union density rate in the government sector is partly explained by differences in labor law that make it easier for public sector employees to organize unions. A second commonly cited reason for the difference is the relative mobility of jobs in the private sector versus the immobility of public or government jobs (i.e., private sector jobs are far more likely to move within the United States or overseas to avoid unionization than are public sector jobs).

In addition to varying over time and across the public and private sectors, union density rates also vary geographically and by industry. States with relatively high union density are located in the Northeast, the Midwest, and the West Coast. In 2008, New York had the highest union density of any state at 24.9%, followed by Hawaii (24.3%) and Alaska (23.5%). Low rates are the norm across the South and in the southwestern states. North Carolina (3.5%), South Carolina (3.7%), and Georgia (3.9%) had the lowest rate of union density in 2008 (Bureau of National Affairs [BNA], 2009). Among industries with high rates of union density in 2008

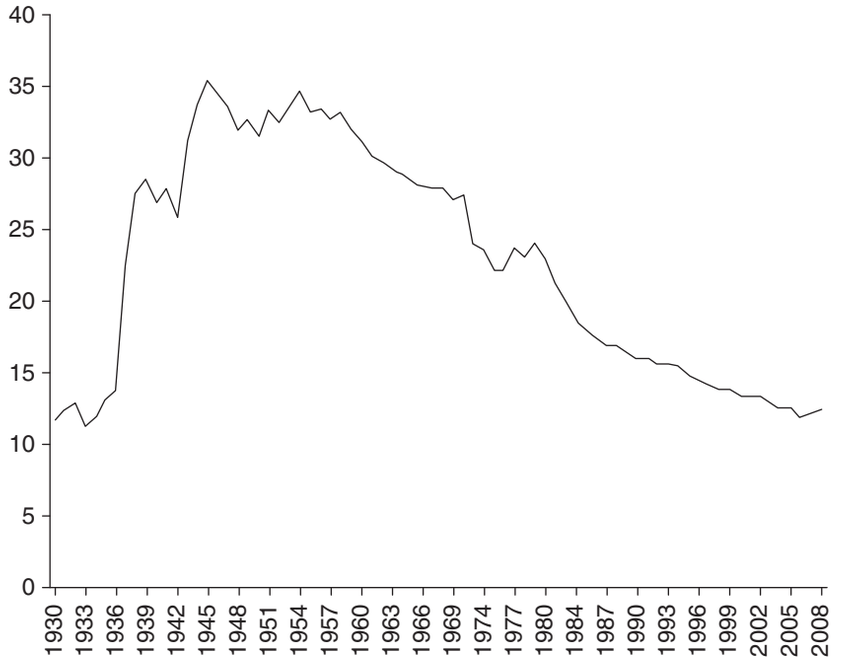


**Figure 16.1** Percentage of Employees That Are Members of a Labor Union in the United States From 1930 to 2008

SOURCES: Data from 1930 to1970 are from Census Bureau Report (2001). Data from 1973 to 1981 are based on results by Hirsch and Macpherson (2009), which uses information from a May CPS data report. Data from 1983 to 2008 were provided directly to the authors from the Bureau of Labor Statistics (BLS, 2009a, 2009b) with the following explanations: (a) Beginning in 2000, data refer to private sector wage and salary workers; data for earlier years refer to private nonagricultural wage and salary workers, (b) Data for 1990 to 1993 have been revised to reflect population controls from the 1990 census, and (c) Beginning in 2000, data reflect population controls from census 2000 and new industry and occupational classification systems. BLS has an online source for the CPS 2000 to 2008 data at http://www.bls.gov/cps/cpslutabs.htm. *Employees* refers to those who receive a wage or salary working in nonagricultural industries in the private and public sectors. No data were available from 1971 to 1972 or for 1982.

were utilities (electric power, natural gas, water supply, etc.) at 26.9%, transportation and warehousing (transportation of passengers and cargo, warehousing and storage of goods, etc.) at 21.3%, and telecommunications (telephone service, cable and satellite television, Internet access, etc.) at 19.3%. Industries with low union density included insurance at 1.0%, finance (banks, savings and loans, brokerages, etc.) at 1.3%, agriculture at 2.8%, food service (restaurants, fast food, caterers, etc.) at 3.1%, and retail trade (stores, catalog sales, etc.) at 5.2% (BNA, 2009a).

While union membership has historically been drawn from the ranks of blue-collar workers in the manufacturing, coal, utilities, transportation, and construction industries, the proportion of the contemporary labor movement's membership made up of white-collar and professional employees has been increasing (although union density for these workers still remains relatively low). Actors, writers, athletes, nurses, teachers, sales representatives, school principals, musicians, and software engineers have joined unions in greater and greater numbers in recent years. The potential for growth in this area presents a great opportunity for American unions. The degree to which they can take advantage of this opportunity may determine how relevant the American labor movement will be in the years ahead.

Labor unions are not by any means solely an American phenomenon. In fact, unions are a significantly more important part of the economies of most industrialized nations than in the United States. In 2006, the United States ranked 22nd of 24 nations, with only France (8.3%) and the Republic of Korea (11.2%) having labor movements that represented a lower percentage of the labor force. Countries with the highest union density rates were Sweden (78%), Finland (74%), and Denmark (70%). It is important to note that union density rates have been declining in most developed countries, although the decline appears to be slower for most other nations than for the United States (Chang & Sorrentino, 1991; Visser, 2006).

## Union Structure and Government

To understand how unions function as organizations, it is necessary to understand how they are structured and governed. The structure and government of most unions is similar to the structure of American government in that it has three levels—local, state or regional, and national.

The basic structural unit of a labor organization is the local union (in some unions the local might be referred to as a lodge or a branch). A local union often consists of the employees in a single workplace (a factory, an office, a warehouse, etc.), although sometimes a local can represent more than one workplace, and in very large workplaces, there can be more than one local union (e.g., one local might represent production and maintenance employees, while a second local represents the office workers). One of the most significant things about a local union is that it is the main point at which members interact with, or experience, the union. The local union is governed based on a local constitution or bylaws that establish how the local will operate. Typically, the local union constitution outlines how officers are elected, what their terms of office will be, what their duties are, as well as how decisions regarding the administration of the local, the bargaining of contracts, and the calling of strikes will be made. The exact structure of a local union will depend in part on its size and on the industry within which it operates. Because most local unions are affiliated with a

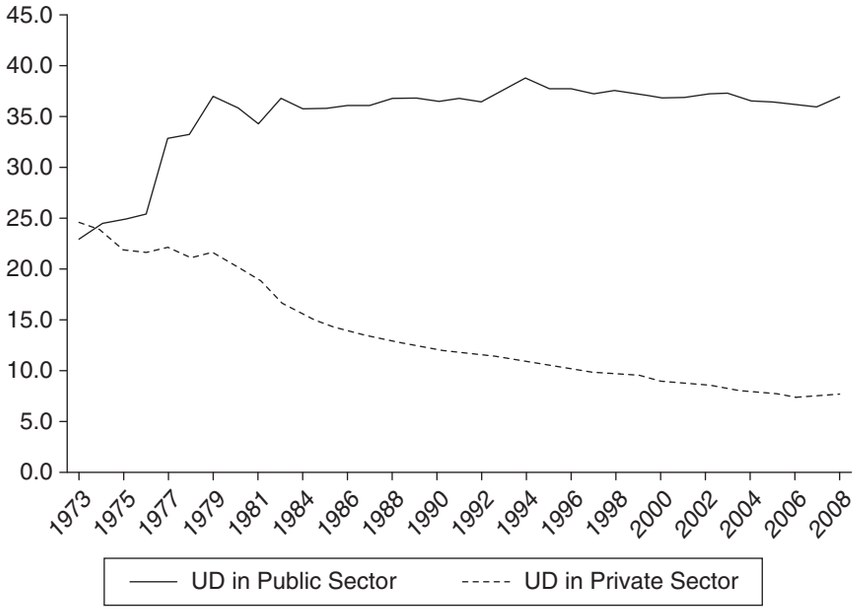


**Figure 16.2** Union Density in the Public and Private Sectors

SOURCES: Data for 1973 to 1981 are from Hirsch and Macpherson (2009). No actual data exist for 1982. Data from 1983 to 2008 were provided directly to the authors by the BLS (2009a, 2009b) with the following explanations: (a) Beginning in 2000, data refer to private sector wage and salary workers; data for earlier years refer to private nonagricultural wage and salary workers, (b) Data for 1990 to 1993 have been revised to reflect population controls from the 1990 Census, and (c) Beginning in 2000, data reflect population controls from Census 2000 and new industry and occupational classification systems.

national or an international union, the local union structure will also be influenced by the practices of the parent organization.

The second level of union structure is the regional or state level. This is an intermediate level of the union that serves as a link between the national and local levels. The regional or state level brings locals belonging to the same national or international union in a given geographic region together for their mutual benefit. This level also often provides professional union representatives to assist the local unions. These individuals usually have significant experience and expertise in organizing, contract negotiations, and grievance handling/arbitration and provide advice and assistance to the local unions.

The national or international union represents the third level of union structure. A national union brings together its local unions within the United States. An international union is similar to a national union except that it has local unions and members in Canada (and sometimes in U.S. territories such as Puerto Rico and Guam). National and international unions perform numerous functions. They bring locals together to use their collective power in negotiations with larger national and multinational employers. National and international unions also provide the formal structure and mechanisms for their members' interests to be voiced and heard in national politics and in addressing other important issues. Additionally, national and international unions support regional/state and local union initiatives and needs by providing services (e.g., legal, research, education) to these levels of the union.

It is important to note that unions—at the local, regional or state, and national/international levels—are mandated by federal law to operate democratically. The Taft-Hartley Act of 1947 (a series of amendments to the NLRA) and the Labor-Management Reporting and Disclosure Act (LMRDA) of 1959, also known as the Landrum-Griffin Act, established basic or minimum requirements for how elections are to be held, who can hold an elected position, and what the terms of office will be. These laws also mandate that all levels of all unions must file detailed reports accounting for every cent that is collected and spent and listing the salaries and expenses of all employees.

There is a fourth level of union structure and government called the *federation.* Unlike the other three levels discussed previously, a labor federation is not a union, nor is it a part of a union. It is a body that brings together many different unions into a loose alliance. This alliance helps the individual unions pursue their common interests by sharing information and bringing the collective authority and resources of the individual unions together. In some ways, a labor federation is analogous to the U.S. Chamber of Commerce in the business community. In the same way that a labor federation is not a union, the Chamber is not a business. Rather, it is a federation of businesses.

There are currently two labor federations in the United States. The first, the American Federation of Labor and Congress of Industrial Organizations (AFL–CIO), was founded in 1955. The second, Change to Win (CTW), was formed in 2005 by unions that disaffiliated from the AFL–CIO. The CTW unions split with the AFL–CIO over a number of policy differences, including CTW's perception that the AFL–CIO was not placing sufficient emphasis on organizing new union members.

In 2009, the AFL–CIO consisted of 56 national and international unions with a total membership of 11 million, while CTW had 7 affiliates totaling 6 million members. While most U.S.-based national and international unions belong to one of these two federations, not all national or international unions are affiliated with a labor federation. In fact, the largest American union, the National Education Association (NEA), does not belong

to either the AFL–CIO or CTW. However, other than the NEA, the unions that not affiliated have relatively small memberships and little influence nationally (Sloane & Whitney, 2010).

## The Industrial Relations Process

### The Three Steps

To understand the role that unions play in labor markets, it is helpful to understand the process by which employees form unions and, once formed, how unions negotiate contracts and resolve day-to-day differences with employers.

The industrial relations, or union–management relations, process consists of three steps. For most private sector workers, these steps are spelled out in the NLRA. For railroad and airline workers, they are spelled out in the Railway Labor Act. For public sector employees who have been granted the rights to organize and bargain and in a few cases to strike, these steps are addressed in various federal and state laws. The processes differ from law to law but generally have much in common.

The first step in the process involves the organization of a union. Under the NLRA, this occurs in one of two ways. If a majority of employees in a workplace sign representation cards, the employees can request that their employer voluntarily recognize the union. If the employer refuses, the employees can then present these cards to the National Labor Relations Board (NLRB) and ask for a certification election (note, while it is necessary to have a majority of employees sign cards to ask for voluntary recognition, only 30% are required to request an election). Once the cards are verified, the NLRB will order that an election be held within 45 to 60 days.

In the time between the filing of cards and the election, the union and the employer will both conduct campaigns to convince the employees to vote for or against unionization. At the end of the campaign, a secret ballot election is typically held. Most employers do not opt to voluntarily recognize a labor union, despite having evidence from a third party that a majority of those workers want to form a union. For the union to be certified, a majority of employees voting in the election must vote for the union. Once a union is certified or voted in, the employer is obliged to engage in bargaining with that union.

The second step in the industrial relations process is collective bargaining. Collective bargaining involves the employer and representatives of the union meeting to negotiate over compensation, benefits, hours, and working conditions for unionized employees. In the bargaining process, the union usually presents proposals for improved wages and benefits, and the employer responds with counterproposals for less generous improvements (or during downturns, reductions in compensation and benefits).

In the small percentage of cases where the parties cannot reach an agreement on the terms of the new collective bargaining contract, one side or the other might engage in strategies to force the other side to change its position. For unions, this would normally take the form of a strike, a tactic in which the union's members refuse to work in an effort to put economic pressure on the employer by disrupting its ability to operate. As a counterbalance to a strike, private sector employers, under certain conditions, can lockout, or voluntarily close, their businesses in an effort to deprive employees of their income.

It is important to note that a very, very small proportion of negotiations involve either a strike or a lockout. The vast majority of collective bargaining negotiations conclude with the two sides agreeing to a contract without any disruption of work. The end result of collective bargaining is a contract (also known as a labor agreement). Contracts are for fixed periods of time, normally 3 to 4 years, although they can be of shorter or longer duration.

Once the parties agree to a contract of a fixed duration, they enter the third step of the industrial relations process. This step is sometimes called the contract administration step. It involves the implementation of a mechanism called a grievance procedure for systematically resolving disputes that arise during the term of the contract. While both parties to the contract may intend to abide by the agreement, there may be incidences when the employer or the union may misinterpret or misunderstand the intent of a specific section in the contract. Disputes might also arise over union concerns that the employer is not following the contract (e.g., an employer does not follow the provisions relating to promotion in choosing a person for a vacated position).

Prior to the passage of the NLRA, unions would sometimes strike over such issues, but today they are settled peacefully through the grievance procedure. This is a quasi-judicial process in which disputes are worked out during a series of meetings involving employer and union representatives. If the issue is not resolved at the first-step meeting, the union can appeal to a second and third step. Each meeting involves different union and management representatives with increasing authority and expertise in addressing the dispute resolution. Most disputes are successfully resolved within the first three steps of the grievance process. However, if no resolution is reached between the parties, the case goes to arbitration, a quasi-judicial process that involves a hearing before an arbitrator (a sort of judge). The arbitrator is usually empowered by both parties to issue a final and binding decision.

A relatively high percentage of cases taken to arbitration involve disciplinary charges against employees (demotions, discharges, etc.). The grievance procedure provides disciplined workers with due process, including an opportunity to present their sides of the story and have their day in court. From the union members' point of view, this is one of the most important benefits a union provides its members—a voice in resolving disputes between an employee and a manager.

### Distinctive Features of the U.S. Industrial Relations System

Just as American unions are somewhat different from their counterparts in other parts of the world, so is the system of industrial relations in the United States. The system is distinctive in four major ways.

#### *Decentralization*

The focal point for collective bargaining in the United States is the individual workplace or bargaining unit (a factory, an office, etc.). This is the unit in which employees organize to form a union. While these bargaining units can be combined into multi-unit structures for the purpose of negotiating a uniform contract across multiple workplaces operated by one employer (as in large corporations like General Motors, U.S. Steel, etc.), the vast majority of collective bargaining in the United States involves a single local union and a single employer negotiating a labor agreement for a single workplace.

While industry-wide collective bargaining is rare in the United States, national labor agreements covering entire industries are much more common in the United Kingdom, western Europe, and Australia. In some U.S. industries, large employers do negotiate company-wide contracts that can cover hundreds of workplaces (this is the case in the auto industry and its three majors employers, General Motors, Ford, and Chrysler), but again, this is the exception and not the rule.

#### *Exclusive Representation*

A second practice that differentiates the U.S. collective bargaining system from others in the world is exclusive representation. The NLRA provides that only one union can represent the employees in a bargaining unit. If workers in a unit decide to organize and are successful in convincing a majority of those employees who vote in the election to support the union, that union becomes the exclusive representative for those employees. No other union can represent any of the workers in that unit. In addition, the union that wins the election acquires the responsibility to represent all employees in that unit, not just the employees who voted for the union.

In many other countries, unions represent only that part of a bargaining unit that chooses to be represented by a particular union. This proportional representation can result in multiple unions representing different percentages of the employees in a given unit (i.e., Union A might represent 40% of the unit; Union B, 30%; Union C, 20%; etc.). Employers are forced to deal simultaneously with multiple unions in that unit, a situation that can be chaotic, confusing, and disruptive.

#### *The Role of Labor Contracts*

In the U.S. system, the heart of any union–management relationship is the contract. These contracts are the end product of bargaining and spell out, with precise language and great detail, what the compensation, benefits, hours, working conditions, local practices, disciplinary processes, and grievance procedures will be. These contracts are often long and complex (the General Motors–United Auto Workers' contract is an almost 1-inch-thick booklet of very small print). This contrasts with the United Kingdom's and western Europe's approach in which negotiations focus on only a handful of issues (wages, some benefits, etc.). The resulting contract is much more limited in scope. Issues not addressed are resolved as they arise, in a more informal process by the parties.

#### *The Role of Government*

A fourth difference between the U.S. system of collective bargaining and the system in most countries is the larger role that the U.S. government plays in determining, regulating, and enforcing the bargaining process. Legislation like the NLRA, the Railway Labor Act, and federal and state public sector labor laws detail how a union is formed, who can be included in that union, the obligations the parties have to bargain with each other, when bargaining occurs, and more.

U.S. labor laws also specify in great detail what the parties must bargain over (mandatory issues); what they cannot bargain over (prohibited issues); and what they may, but are not required to, bargain over (voluntary issues). Again, in the United Kingdom and western Europe, the government does not get involved to as great a degree in either the process or the substance of negotiations. These issues are largely left to the parties.

## The Impact of Unions on Labor Markets

As suggested earlier, economists have traditionally viewed unions as monopolies that interfere with the efficient operation of labor markets. They do so in three main ways.

First, unions raise wages above the competitive levels set by the market. In doing so, they cause employers to hire fewer workers than they would otherwise. The evidence does, in fact, indicate that unions increase wages. In 2008, the BLS (2009c) reported that the median weekly earnings of union members employed full time was $886, while full-time nonunion workers earned $691, a difference of 28% (see Figure 16.3).

A comparison of earnings by occupation indicates that this earnings premium is present in most occupations (see Table 16.1).

The impact of unions on wages is even greater than indicated by the union–nonunion wage differential. This is because nonunion employers often raise the wages of their employees above what they would otherwise pay to reduce the likelihood that their employees will organize a union. The phenomenon of unions indirectly causing an increase

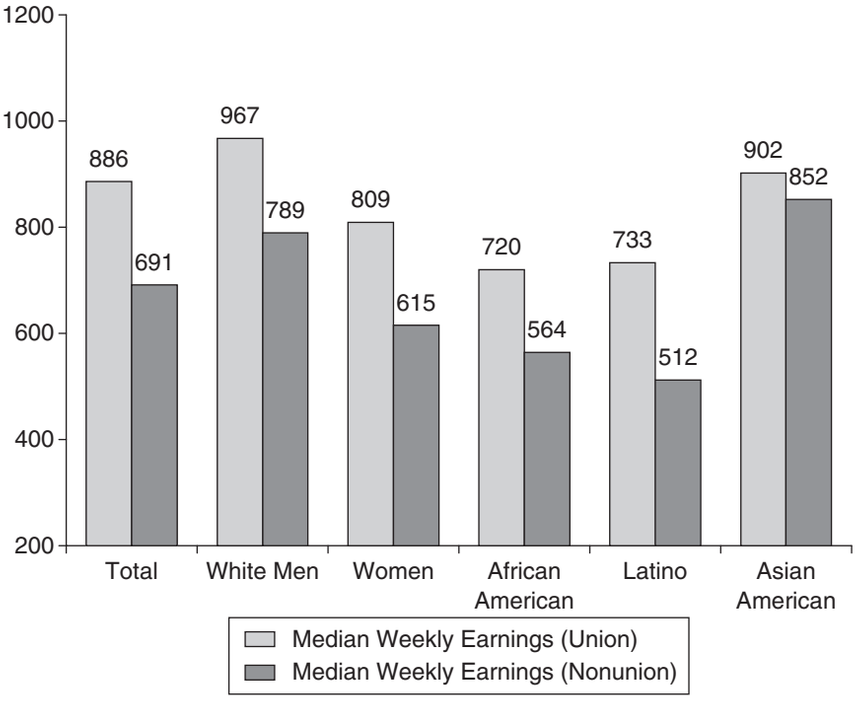


**Figure 16.3** Comparison of Union and Nonunion Median Weekly Earnings 2008
SOURCE: BLS (2009c).

in the wages of nonunion employees is called the *union threat effect* (Filer, Hamermesh, & Rees, 1996).

Economic theory suggests that artificially raising wages results in a reduction in the number of jobs because employers buy less labor as the price increases. And as those employees who have lost their jobs search for work in the nonunion sector, they bid down wages there. In essence then, some of the wage increases won by unions come at the expense of lower-paid or unemployed workers. For this reason, many economists argue that the interference of unions in the operation of the market causes inefficiencies. This leads to the conclusion that, on balance, unions play a harmful role in the market.

Many economists believe that unions contribute to inefficiency in a second way by engaging in strikes. When unions call strikes in an effort to increase their bargaining power, productivity falls. At the firm level, this ultimately lowers profitability; at the national level, it reduces gross national product (GNP) (Freeman & Medoff, 1984).

Third, economists believe that unions reduce efficiency through the imposition of work rules and work restrictions with which nonunion employers are not saddled. One often-cited example is work jurisdiction rules in which employees have strictly observed job descriptions that prevent them from doing even the simplest of tasks that are not a part of their jobs. Where work jurisdiction rules exist, some economists argue that these provisions have a negative impact on individual and firm-level productivity (Filer et al., 1996).

However, other economists question this view of unions. They believe that "markets are competitive enough to give unions little or no power to extract monopoly wages" (Freeman & Medoff, 1984, p. 7). They reason that unions really acquire monopoly power only when they organize an entire market or are present in a noncompetitive market (Filer et al., 1996). They also point out that strikes involving 1,000 workers or more have fallen steadily in this country from a modern high of 470 in 1952 to 44 in 1990 to 15 in 2008 and that the percentage of estimated working time lost for those same years fell from 0.14 to 0.02 to 0.01 (BNA, 2009). They also note that strikes occur in other industrialized countries and usually at higher rates that more than offset the cost of strikes in the United States. Last, they argue that restrictive work rules may have been a problem in the past but that many unions have worked closely with employers to eliminate these practices and to find ways to more efficiently produce goods or provide services.

Over the last 25 years, another perspective on unions, based on the writings of Hirschman (2007), but most effectively articulated by Freeman and Medoff (1984), has gotten significant attention and has influenced economists' views of unions in important ways. This collective voice/institutional response perspective suggests an alternative to the classic market mechanism of exit and entry. Economic theory posits that exit occurs in a perfectly competitive market in which "no individual can be made better off without making someone worse off " (Freeman & Medoff, p. 8). The freedom that dissatisfied employees have to leave bad employers and go to work for good ones contributes to the efficiency of the market. If this system works as assumed, unions can interfere only with the free operation of the market and create inefficiencies (Bennett & Kaufman, 2007).

Freeman and Medoff (1984) contend that employees can deal with workplace problems in two ways. They can exit (quit—the only option available to workers according to classical economic theory) or they can engage in voice by speaking up and trying to change the conditions with which they are dissatisfied. Regarding the exit option for dissatisfied workers, Freeman and Medoff argue that these self-regulating mechanisms of markets are not perfect because actors do not have complete information and there are significant mobility costs for employees who choose to exit. And should employees consistently exit the firm when dissatisfied, the

**Table 16.1** Earnings by Occupation, 2008 Full-Time Wage and Salary Workers' Median Weekly Earnings in Dollars

| | *Union* | *Nonunion* | *Percentage of Union Advantage* |
|---|---|---|---|
| Community and social services occupations | 983 | 743 | 32.3 |
| Education, training, and library occupations | 974 | 765 | 27.3 |
| Arts, design, entertainment, sports, and media occupations | 1,110 | 858 | 29.4 |
| Health care practitioner and technical occupations | 1,070 | 943 | 13.5 |
| Health care support occupations | 526 | 457 | 15.1 |
| Food preparation and serving-related occupations | 502 | 398 | 26.1 |
| Building and grounds cleaning and maintenance occupations | 596 | 412 | 44.7 |
| Construction and extraction occupations | 992 | 621 | 53.0 |
| Installation, maintenance, and repair occupations | 1,002 | 729 | 37.4 |
| Production occupations | 765 | 567 | 34.9 |
| Transportation and material moving occupations | 789 | 550 | 43.5 |

SOURCE: BLS (2009c).

result is high turnover in the workforce, which is costly to employers. Voice can, in fact, reduce costs to both employers and employees by drawing attention to bad employment practices and resolving them. Freeman and Medoff point out that voice expressed by an individual is much less likely to be effective in resolving problems, particularly in a large workforce, than is voice expressed by a group. And because the most effective way to express collective voice is through a union, they conclude that unions can have a positive impact on efficiency in a workplace.

The view that unions interfere with the efficient operation of the market continues to be the prevailing opinion in the field of economics. But the case made by Freeman and Medoff (1984) that, in addition to their negative monopoly face, unions also have a positive voice face has forced the economics profession to reconsider the conventional wisdom about unions.

## Conclusion

By any account, labor unions have played a significant role in American society for most of the nation's history. And while their influence has declined over the last 25 years, the modern American labor movement represents millions of employees in thousands of large and small workplaces across the country.

Most economists have viewed unions through the lens of neoclassical economic theory, concluding that they act as monopolies that create inefficiencies in the labor market, resulting in the loss of jobs and greater income inequality in the workforce. In their view, unions have also had a negative impact on efficiency through the conduct of strikes and by the institution of cumbersome work rules and work restrictions.

This assessment of unions has been challenged in recent years by a minority of economists who downplay the monopoly face of unions. These scholars argue that strikes no longer cause significant disruption to the economy and that unions have greatly loosened restrictions on work rules. And they argue that unions have a second voice face that plays a positive role in the workplace by allowing employees to address problems that would otherwise cause them to exit or quit. This side of unions benefits both employees and employers by reducing turnover, improving productivity, and bringing fair treatment and due process, two of the core values of American democracy, to the workplace.

It remains to be seen whether this minority view of unions can make inroads into the traditional view that has prevailed in the field of economics for a very long time.

## References and Further Readings

Bamber, G. J., Lansbury, R. D., & Wailes, N. (Eds.). (2004). *International and comparative employment relations* (4th ed.). London: Sage.

Bennett, J. T., & Kaufman, B. S. (Eds.). (2007). *What do unions do? A twenty-year perspective.* Piscataway, NJ: Transaction Publishers.

Bok, D. C., & Dunlop, J. T. (1970). *Labor and the American community.* New York: Simon & Schuster.

Booth, A. (1995). *The economics of trade unions.* New York: Cambridge University Press.

Borjas, G. (2009). *Labor economics* (5th ed.). New York: McGraw-Hill.

Bureau of Labor Statistics. (2009a, February 24). [Percent of wage and salary workers who were union or employee association members by sector, annual averages 1983–2008]. Unpublished raw data.

Bureau of Labor Statistics. (2009b, February 3). [Union affiliation of employed wage and salary workers by sex, race, and Hispanic or Latino ethnicity, annual averages 1983–2008.] Unpublished raw data.

Bureau of Labor Statistics. (2009c). *Union members in 2008.* Retrieved March 31, 2009, from http://www.bls.gov/news.release/pdf/union2.pdf

Bureau of National Affairs. (2009). *Work stoppages involving 1,000 or more workers, 1947–2008.* Retrieved March 31, 2009, from http://www.bls.gov/news.release/wkstp.t01.htm

Census Bureau Report. (2001). Chapter D: Labor. In *Historical statistics of the United States: Colonial times to 1970* (pp. 121–182). Retrieved March 31, 2009, from http://www2.census.gov/prod2/statcomp/documents/CT1970p1-05.pdf

Chaison, G. (2006). *Unions in America.* Thousand Oaks, CA: Sage.

Chang, C., & Sorrentino, C. (1991, December). Union membership statistics in 12 countries. *Monthly Labor Review, 114*(12), 46–53.

Dine, P. (2007). *State of the unions: How labor can strengthen the middle class, improve our economy, and regain political influence.* New York: McGraw-Hill.

Dubofsky, M. (2000). *We shall be all: A history of the Industrial Workers of the World.* Urbana-Champaign: University of Illinois Press.

Dunlop, J. T. (1990). *The management of labor unions: Decision making with historical constraints.* Lexington, MA: Lexington Books.

Eaton, A., & Keefe, J. H. (Eds.). (2000). *Employment dispute resolution and worker rights in the changing workplace.* Ithaca, NY: Cornell University Press.

Ehrenberg, R. G., & Smith, R. S. (2008). *Modern labor economics: Theory and public policy* (10th ed.). New York: Addison-Wesley.

Filer, R., Hamermesh, D., & Rees, A. (1996). *The economics of work and pay* (6th ed.). New York: HarperCollins.

Fletcher, B., & Gapasin, F. (2008). *Solidarity divided: The crisis in organized labor and a new path toward social justice.* Berkeley: University of California Press.

Freeman, R., & Medoff, J. (1984). *What do unions do?* New York: Basic Books.

Freeman, R. B., & Rogers, J. (2004). *What workers want.* Ithaca, NY: Cornell University Press/ILR Press.

Friedman, M., & Friedman, R. (1962). *Capitalism and freedom.* Chicago: University of Chicago Press.

Friedman, M., & Friedman, R. (1980). *Free to choose.* New York: Harcourt Brace Jovanovich.

Gall, G. (Ed.). (2006). *Union recognition: Organizing and bargaining outcomes.* London: Routledge.

Gould, W. (2004). *A primer on American labor law.* Cambridge: MIT Press.

Graham, I. (2003). It pays to be union, U.S. figures show. *Labour Education Online, 128,* 13–16. Retrieved March 31, 2009, from http://www.ilo.org/public/english/dialogue/actrav/publ/128/3.pdf

Greenhouse, S. (2009). *The big squeeze: Tough times for the American worker.* New York: Anchor.

Hirsch, B., & Addison, J. (1986). The economic effects of employment regulations: What are the limits? In B. Kaufman (Ed.), *Government regulation of the employment relationship* (pp. 125–178). Madison, WI: Industrial Relations Research Association.

Hirsch, B., & Macpherson, D. A. (2009). *U.S. historical tables: Union membership, coverage, density, and employment 1973–2008.* Retrieved March 31, 2009, from http://www.unionstats.com/

Hirschman, A. O. (2007). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states.* Cambridge, MA: Harvard University Press.

Hyman, R. (2001). *Understanding European trade unionism.* London: Sage.

Kochan, T., & Lipsky, D. (Eds.). (2003). *Negotiations and change.* Ithaca, NY: Cornell University Press.

Lewin, D., & Peterson, R. (1988). *The modern grievance procedure in the United States.* Westport, CT: Quorum.

Lipset, S. M. (1997). *American exceptionalism: A double-edged sword.* New York: W. W. Norton.

Mort, J. (Ed.). (1998). *Not your father's union movement: Inside the AFL-CIO.* New York: Verso.

Murolo, P., Chitty, A. B., & Sacco, J. (2003). *From the folks who brought you the weekend: A short, illustrated history of labor in the United States.* New York: New Press.

Reynolds, L. G., & Taft, C. H. (1956). *The evolution of the wage structure.* New Haven, CT: Yale University Press.

Simons, H. C. (1948). *Economic policy for a free society.* Chicago: University of Chicago Press.

Sloane, A. A., & Whitney, F. (2010). *Labor relations* (13th ed.). Upper Saddle River, NJ: Prentice Hall.

Taft, P. (1954). *The structure and government of labor unions.* Cambridge, MA: Harvard University Press.

Visser, J. (2006). Union membership statistics in 24 countries. *Monthly Labor Review, 129*(1), 38–49.

Yates, M. (1998). *Why unions matter.* New York: Monthly Review Press.

# 17

# Game Theory

Indrajit Ray
*University of Birmingham*

Game theory, within the discipline of economics, has perhaps been the most discussed field in the past half century—it has revolutionized the study of economics, inspired hit films, and crept into the popular awareness. The success story of game theory as a research area can easily be found in the advancement of field-specific journals and the many Nobel Prizes awarded to game theorists within the past two decades.

As it offers powerful tools for understanding strategic interactions, game theory is being increasingly applied not just within economics (such as industrial organization, political economy, international trade, environmental economics) but also in other subject areas, such as politics, sociology, law, biology, and computer science. It is not surprising that game theory has gained enormous popularity and importance as a teaching module within any (undergraduate or graduate) program in economics and other related subject areas.

So what exactly is game theory? What is a game for that matter; what makes a game, a game that's being theorized here? The answer is simple: Game theory is nothing but an interactive decision theory involving more than one individual (decision maker). Game theory offers mathematical models to study conflict and cooperation between intelligent and rational decision makers. The decision makers here are called the players. These players are assumed to be intelligent and rational in the sense that they can identify their own objectives and can choose the best among a set of alternatives by comparing outcomes and performing necessary calculations.

There are many real-life examples of games—board games such as chess and checkers are games that are theorized. However, many televised "game" shows are not games as they do not involve interactive decision making. Indeed, many real-life situations can be modeled and analyzed using game theory. Common examples include auctions of goods ranging from artwork to 3G mobile network license, wage bargaining by any union, voting in committees, and so on.

Game theory, as a subject and a field of research, has evolved over the decades in the past century. In the early years, game theory clearly was identified as an applied topic in mathematics (game theorists then were applied mathematicians, unlike present-age game theorists who are pure economic theorists). The subject perhaps got its independent status in the 1940s with the publication of Von Neumann and Morgenstern's (1944) book and the doctoral thesis of John Nash, later published as papers (Nash, 1950a, 1950b).

Broadly speaking, the theory can be classified under two areas: noncooperative game theory and cooperative game theory. The models and solutions used to analyze conflict (and possible cooperation) among interactive individuals (decision makers) constitute noncooperative theory. On the other hand, models of cooperative game theory assume all possible coalitions may form and thus take coalitions (and not individuals) as the building blocks of the theory. The solutions of the cooperative theory analyze possible ways of sharing the benefits achieved by cooperation in a coalition among individuals.

Research in game theory in the past 60 years or so has made sharp turns in terms of its focus. In its early years (in the early twentieth century) as a topic in applied mathematics, a specific form of a two-person noncooperative game, called the "zero-sum game" (in which one player wins what the other player loses), was the center of attention. Then, in the 1950s and the 1960s, the study of cooperative games was in fashion while from the 1970s onwards, noncooperative theory stole the focus. Perhaps it is an irony that John Nash, who essentially initiated the

modern theory of noncooperative games with the concept of the Nash equilibrium (Nash, 1950b), was also the author of a path-breaking work in cooperative theory known as the Nash bargaining solution (Nash, 1950a).

Following the latest trend in the subject of game theory, we mainly deal with the modern notions of noncooperative game theory (with almost no mention of the early research on zero-sum games) in this chapter. Only the last section of this chapter is devoted to the field of cooperative game theory. We believe this treatment is justifiable in a reference collection for twenty-first-century economics!

## Models

Models of noncooperative games are nothing but tools to turn any appropriate real situation into a "game." The modeler, before analyzing the game, must translate any situation into the language of the theory. The model describes a well-defined problem or a situation as a game. It is important to note that the model, once well formulated, stands independent of the analysis and the proposed "solutions" of the corresponding game. The theorists and the practitioners may differ in their analysis of the game; however, the model should be objectively presented. In this section, we will present different models of noncooperative game theory that are suited to describe different situations as games. No analysis or solution of the games will be presented in this section; the following section is devoted to such analyses.

A noncooperative game is used to describe any strategic interaction among agents and is characterized by who the "players" are, what the possible alternative "actions" or "strategies" available to each player are, and what the "payoffs" to each player are when different strategy combinations are chosen. The key ingredient of a game is the fact that the payoff to a player depends on not just his or her own choice of strategy or action but on other players' strategies as well and that every player is fully aware of this interdependence.

Noncooperative games can be broadly classified into two groups depending on how they are played. Some games are played simultaneously; that is, in these games, the players choose their strategies simultaneously, not observing (any part of) each others' choices. Note that the phrase "simultaneously played" is not used literally; it need not imply that the choices are made at the same time, as long as the players do not observe others' strategies. For example, voting can be viewed as a simultaneously played game, played by the voters who can very well cast their votes at different times.

In contrast with simultaneously played games, some games are played sequentially over time and thus are called sequentially played games. The players in these games choose their actions in a particular sequence according to the rules of the game.

Simultaneously played games are often called normal-form or strategic-form games. The ingredients of any simultaneously played game are the players, their possible actions, and the outcomes depending on the chosen actions.

We focus on "finite" games where the set of players denoted by $N = \{1, \ldots, n\}$ of the game is a finite set and the sets of actions (often called the "pure strategies") available to each player, denoted respectively by $S_1, \ldots, S_n$, are all finite. $s_i \in S_i$ denotes a typical action or a pure strategy for any player *i.* The outcomes of the game depend on the choices made by all the players. An outcome can thus be associated with an $n$-vector of actions (pure strategies) denoted by $s = (s_1, \ldots, s_n) \in S = \prod_i S_i$ (the product set). Such a vector $s \in S$ is known as a strategy profile indicating a chosen strategy by each of the players. We use $s_{-i}$ to denote the choices made by all but player *i.*

Each player has a preference ordering over these outcomes. The preferences can be represented by utility functions denoted respectively by $U_1, \ldots, U_n$. These provide payoff numbers to the players for each outcome of the game. Such a game, $G = < N, S_1, \ldots, S_n, U_1, \ldots, U_n >$ provides a complete description of the situation and the rules of the game.

It is important to highlight here that we assume that the rules of the game and the situation described in the model are known to the players. Moreover, the players know that the other players know and that the other players know that they know and so on. In other words, we assume that the game is "common knowledge" among all players. It is indeed important to maintain this assumption in all the models we discuss here as we need our players to realize that they all play the same game.

A two-person strategic-form game can easily be described in a table, often called the normal-form table. It is a convention to represent the rows of the table as the pure strategies of Player 1 (thus called the row player) and the columns as those of Player 2 (thus called the column player). Each cell of the table is thus a strategy combination of the game and therefore represents an outcome that is associated with two payoff numbers. The first number in each cell is Player 1's payoff while the second is for Player 2.

We now consider some examples of simultaneously played games in which there are only two players, each with two strategies. All these games thus can be represented by a $2 \times 2$ normal-form table.

Our first $2 \times 2$ noncooperative game has the feature that in each of its outcomes, a player wins the amount that the other loses. Such two-person noncooperative win-lose games are known as zero-sum games as the sum of the payoffs in each cell is zero. The game can be associated with a situation of "matching pennies" where both players put a penny on the table with head (*H*) or tail (*T*) up. Player 1 wins (Player 2 loses) when both pennies have the same side (*HH* or *TT*) while Player 2 wins (Player 1 loses) otherwise. The table in Figure 17.1 describes the game.

The second game is perhaps the most talked about game in economics. The story behind the game involves two convicts who may or may not admit their guilt individually; the game, however, applies to many social situations in general. The two choices available to the players are to cooperate or not to cooperate (defect) with the fellow player. As the table in Figure 17.2 describes, the outcome in which both players choose to cooperate is socially optimal, in the sense that it maximizes the sum of the payoffs achieved from different outcomes; however, as we will find out later, this outcome may not be achievable when the players play this game just once (repeating the same game will lead to the cooperative outcome).

The following two games describe situations involving desirable coordination between two players. The first situation (as in Figure 17.3) is that of pure coordination between two choices *A* and *B*, where the outcomes are ranked in the same order by both players. Both players will get positive payoffs if they coordinate to play the same strategy and get nothing if they do not; moreover, playing *A* leads to a better outcome to both than that from playing *B*.

The second situation (as in Figure 17.4) is often called the battle of the sexes, where the two players, the husband and the wife, choose to go to either a football match (*F*) or a concert (*C*). This game also has the coordination feature that the players will get positive payoffs if they play the same strategy and get nothing if they do not. However, the two coordinated outcomes are ranked differently by the players; the husband prefers the outcome (*F*, *F*) while the wife prefers (*C*, *C*).

The final example in this series, called "chicken," describes a situation where each of the two players can be aggressive (*A*) or passive (*P*), resulting in the outcomes depicted in Figure 17.5.

Tables are useful to describe games of two players only. Clearly, normal-form games may involve more than two players. Tables may still be used to describe three-person games, in a manner suggested below; however, for games with more than three players, one cannot use the tabular representation.

Consider a three-person game in which each player chooses either *S* or *N*, as described in Figure 17.6. The strategies of Player 1 (the row player) are the two rows, those of Player 2 (the column player) are the two columns, and those of Player 3 are the two matrices. There are eight possible outcomes corresponding to the eight ($2 \times 2 \times 2 = 8$; each of three players has two strategies) strategy profiles. These are the eight cells in the two tables.

The payoffs from the outcomes are given in the respective cells with the three numbers, respectively, to Players 1, 2, and 3. For example, when Player 1 plays *S*, Player 2 plays *N*, and Player 3 plays *S*, they get payoffs of −1, −2, and −1, respectively.

| | *H* | *T* |
|---|---|---|
| *H* | 1, –1 | –1, 1 |
| *T* | –1, 1 | 1, –1 |

**Figure 17.1** Matching Pennies

| | *Cooperate* | *Defect* |
|---|---|---|
| *Cooperate* | 3, 3 | 0, 4 |
| *Defect* | 4, 0 | 1, 1 |

**Figure 17.2** Prisoner's Dilemma

| | *A* | *B* |
|---|---|---|
| *A* | 2, 2 | 0, 0 |
| *B* | 0, 0 | 1, 1 |

**Figure 17.3** Coordination Game

| | *F* | *C* |
|---|---|---|
| *F* | 2, 1 | 0, 0 |
| *C* | 0, 0 | 1, 2 |

**Figure 17.4** Battle of the Sexes

| | *A* | *P* |
|---|---|---|
| *A* | 0, 0 | 7, 2 |
| *P* | 2, 7 | 6, 6 |

**Figure 17.5** Chicken

| | | *S* | *N* |
|---|---|---|---|
| | | *S* | *N* |
| *S* | *S* | 1, 1, 1 | –1, –2, –1 |
| *N* | *S* | –2, –1, –1 | 1, 1, 1 |

| | | *N* | |
|---|---|---|---|
| | | *S* | *N* |
| *S* | *S* | –1, –1, –2 | 1, 1, 1 |
| *N* | *S* | 1, 1, 1 | 0, 0, 0 |

**Figure 17.6** A Three-Person Game

The second class of models in noncooperative theory is known as sequentially played games. As the name suggests, these games are used to describe situations in which players take decisions in a specified sequence as in board games. These games are also known as extensive-form games and are described by a structure called a game tree.

As in normal-form games, the (main) ingredients of an extensive-form game are players, their choices, the outcomes (depending on the choices made), and the payoffs to the players from each outcome. All these are described in a game tree with nodes and branches.

An extensive-form game can be identified as a rooted tree with a set of nodes, one of which is a distinct initial node. All the nodes other than the initial one are connected by a precedence function; that is, for any node, there is an immediate predecessor of that node. The terminal nodes are those for which there exists no successor.

A path is a finite sequence of nodes from the initial to a terminal node. Note that in these games trees, any path, starting from the initial node and ending at a terminal node, can be recognized by the particular terminal node at which it ends. In the next section, we will see that this important feature helps us to solve such games easily.

Formally, the set of players is $N = \{1, 2, \ldots, n\}$. All nonterminal nodes are partitioned into $n$ subsets, $X_1, \ldots, X_n$, where $X_i$ is the set of nodes at which player $i$ moves. At any $x \in X_i$, player $i$ has to choose an action that is identified by an immediate successor node in the tree. For game trees that are finite (i.e., the game ends after finitely many moves), we can observe, at the completion of the game, a path leading to a terminal node. Each terminal node therefore can be associated with an outcome of the game and will be assigned with payoff numbers for each player.

Consider, for example, the following version of an "ultimatum game" of splitting \$10 in which Player 1 moves first and has three choices of keeping ($K$) the sum to himself or herself, sharing ($S$) it equally with the other player, or giving ($G$) the amount to the other player. Player 2 then moves, after observing Player 1's choice, and has two choices of accepting ($A$) or rejecting ($R$) Player 1's call. If Player 2 accepts the proposal, he or she gets the corresponding outcomes; otherwise, he or she gets nothing. The tree in Figure 17.7 describes the game in extensive form. The payoffs are at the end of each outcome (terminal node) with the first number to Player 1 and second to Player 2.

Extensive-form games differ from normal-form games not just in structure but also rather crucially in the issue of what information is available to (some of) the players as the game progresses. The information structure becomes a component of the model and is described suitably in the tree.

Some players may receive information during the game as a part of the underlying situation. Information given to the players may differ depending on the situation. The different information may come to some of the players from "nature" (natural course), denoted by Player 0 in the game, or from actions chosen by other players.

Such informational uncertainties during the game are denoted by information sets. An information set is a set of nodes in which a player who is making a choice has no information to distinguish the nodes in that set while other players may be able to do so.

Consider, for example, a card game in which Player 1 first picks a card from a deck and sees its color (that Player 2 cannot). The color of the card, which may be black ($B$) or red ($R$) with equal chances, is thus the information given to Player 1. After seeing the color, Player 1 can "raise" ($r$) the stake or "fold" ($f$). Player 1 wins if the color of the card, when revealed, is red. If Player 1 raises, Player 2, who does not know the color, can then "meet" ($m$) or "pass" ($p$).

The tree in Figure 17.8 describes the situation in which first nature (also called "Player 0") chooses the color of the card (black or red with probability ½ each) that Player 1 observes before making a choice. The two nodes where Player 2 has to move are put under one set. These two nodes are not distinguishable from the viewpoint of Player 2, who in both cases has only observed that Player 1 has raised (chosen $r$); however, Player 2 does not have the information regarding the color of the card (whether it is $B$ or $R$) and therefore cannot distinguish between the nodes. (The tree here is used simply to describe the structure, and the payoffs at the end are omitted as the stakes in the card game can be anything.)

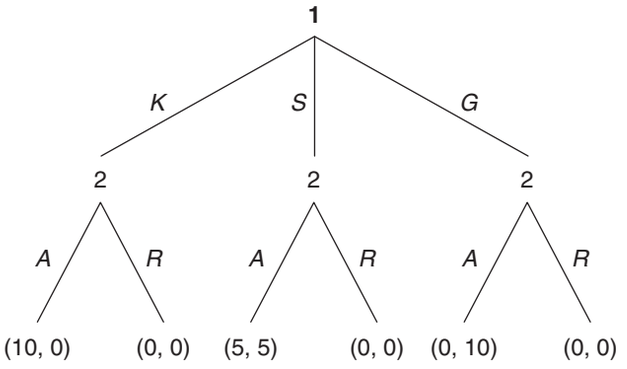


**Figure 17.7** An Ultimatum Game

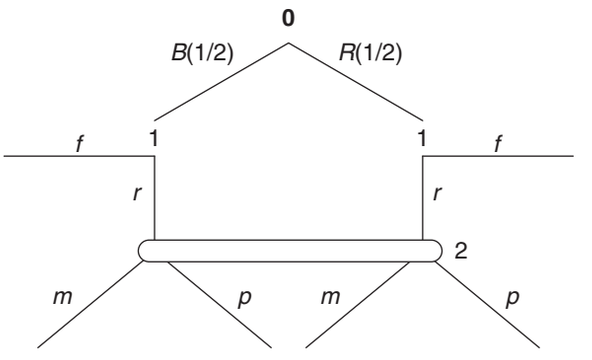


**Figure 17.8** A Card Game

Consider also a situation in which Players 1 and 2 must decide whether to carry an umbrella when leaving home. They know that there is a 50–50 chance of rain. Each player's payoff is −5 whether he or she does not carry an umbrella and it rains, −2 if he or she carries an umbrella and it rains, −1 if he or she carries an umbrella and it is sunny, and 1 if he or she does not carry an umbrella and it is sunny. Player 1 learns the weather before leaving home; Player 2 does not, but he or she can observe Player 1's action before choosing his or her own.

The above situation can be described by the tree in Figure 17.9 where nature (Player 0) first chooses the weather condition—rain (*R*) or sun (*S*)—with equal chances (probability 1/2). Player 1 learns the condition, and in each of the two cases (nodes), he or she has two choices: either to carry an umbrella (*U*) or not to carry (*N*). Player 2 then observes Player 1's choice (but still does not know the precise weather condition) and then must choose either to carry an umbrella (*U*) or not to carry (*N*).

The tree in Figure 17.9 has four nodes in which Player 2 makes a choice. These four nodes are put into two different information sets for Player 2, who does not know the weather condition but can observe Player 1's move. Player 2 observes Player 1 carrying an umbrella in the two nodes in the left information set in the tree; however, he or she cannot distinguish these nodes as he or she does not learn the weather condition. Similarly, in the right information set, he or she sees Player 1 without an umbrella and knows nothing about the weather. As earlier, the payoffs are given at each of the terminal nodes.

Both the examples above describe situations in which (some of the) players get some information (such as the color of the card in the card game) during the game. Thus, when the game is played, players will have different information that is indicated by players' information sets. These games are called games with imperfect information.

We should make a comment here that games with imperfect information are modeled distinctly from situations (games) where players may have different information that is relevant to the game, before the game starts. This sort of prior (to the game) and private (to individual players) information is modeled as games with incomplete information and will be discussed later in this section.

Games that have no informational asymmetry among the players at any stage (prior to or during the game) are known as games with complete and perfect information (such as board games like chess).

Before discussing games with imperfect information further, we should spell out an important assumption behind these models. We assume that the players remember all the past moves and the information provided to them in any stage of the game. Hence, these (extensive-form) games are called games with "perfect recall."

Extensive-form games with imperfect information distinguish players' (pure) strategies from their chosen actions. Actions are players' moves or choices that they decide upon whenever they have to during the game. Strategies, on the other hand, are plans for the whole game. In an extensive-form game, players may move in different time periods, in different stages (nodes) in the game within different information sets. The strategies therefore should be "complete" plans, in the sense of being exhaustive over time stages and being contingent upon the information sets; that is, a strategy for a player must describe what the player does whenever he or she has to move (at every stage and at each of his or her information sets in the game).

For example, in the above card game, Player 1 has to move in two information sets, each being a singleton (with only one node). In the first node (singleton information set), Player 1 knows that the color of the card is black, while in the second node (singleton information set), the color is red. The action to be chosen in each node is either *r* or *f*. A strategy for Player 1 must completely specify what he or she does when the color is black and what he or she does when the color is red. *rf* denotes such a strategy, which implies that Player 1 chooses *r* (raises) when the color is black and chooses *f* (folds) when the color is red. There are thus four such possible strategies—namely, *rr*, *rf*, *fr*, and *ff*—that describe completely what Player 1 does contingent on the information on the color of the card.

Similarly, in the umbrella game, the strategies of Players 1 and 2 should completely specify what they do in the different information sets in which they have to choose an action. In this game, Player 1 has two information sets,

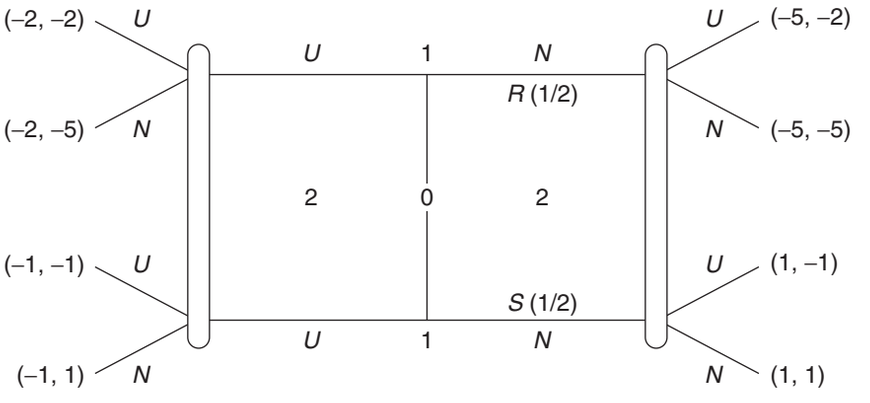


**Figure 17.9** The Umbrella Game

and both are singletons (consisting of only one node), as he or she knows the weather condition. Player 2 also has two information sets, as explained earlier, each with two nondistinguishable nodes, as he or she observes Player 1's choice but does not know the weather condition.

Player 1's (pure) strategy should specify what he or she plans to do when it rains and what he or she does when it is sunny. For example, the plan *UN* denotes that Player 1 chooses to carry an umbrella (*U*) when it rains and chooses not to carry (*N*) when it is sunny. The four possible pure strategies for Player 1 thus are *UU*, *UN*, *NU*, and *NN* denoting what he or she does when it rains and when it does not.

Similarly, the strategies of Player 2 specify what actions he or she chooses in two different information sets he or she has to act on. A complete plan for Player 2 thus also has two components; however, it is contingent on his or her information only, that is, his or her observation of Player 1's move (and not the weather condition). For example, the plan for Player 2, denoted by *UN*, suggests that Player 2 chooses to carry an umbrella (*U*) in his or her left information set when Player 1 carries an umbrella and chooses not to carry (*N*) in his or her right information set when Player 1 does not carry. Thus, the four possible pure strategies for Player 2 also are *UU*, *UN*, *NU*, and *NN;* however, these denote what Player 2 does when Player 1 carries an umbrella and when Player 1 does not.

It is sometimes easier to analyze normal-form games. Given the strategies or complete plan of actions, one can now transform an extensive-form game into a normal-form game. Such normal-form representations include the players and their complete plans as strategies.

For example, the normal form corresponding to the umbrella game has four (pure) strategies for each player. It can be described by a table as shown in Figure 17.10. Each cell in the table denotes an outcome corresponding to the profile of plans chosen by the players with payoffs being the expected payoffs from following those plans.

To understand the (expected) payoffs illustrated in each cell, take, for example, the case where the players follow the plans *UN* and *UN*, respectively. Under these plans, if it rains, Player 1 will carry the umbrella (as he or she knows that it will rain) and Player 2 will also carry (as he or she sees Player 1 carrying), while if it is sunny, Player 1 will not carry the umbrella (as he or she knows the weather) and Player 2 will also not carry (as he or she sees Player 1 not carrying). The game thus ends up in either of the two terminal nodes where the payoffs are (–2, –2) and (1, 1). With a 50–50 chance of rain, the expected payoffs thus will be (–0.5, –0.5).

As mentioned earlier, players may have private information before they start playing the game, which is not in the description of the game. This sort of prior and private information can be modeled as types of players. These games are known as games with incomplete information or Bayesian games.

| | *UU* | *UN* | *NU* | *NN* |
|---|---|---|---|---|
| *UU* | –1.5, –1.5 | –1.5, –1.5 | –1.5, –2 | –1.5, –2 |
| *UN* | –0.5, –1.5 | –0.5, –0.5 | –0.5, –3 | –0.5, –2 |
| *UN* | –3, –1.5 | –3, –3 | –3, –0.5 | –3, –2 |
| *NN* | –2, –1.5 | –2, –2 | –2, –1.5 | –2, –2 |

**Figure 17.10** Normal-Form Table for the Umbrella Game

A Bayesian game is described by the set of players, the strategies of players, the different types of the players, prior beliefs held by the players over others' types, utility functions depending on the type, and strategy profiles.

For example, in the Bayesian game described by the tables in Figure 17.11, each player has two strategies, and Player 2 has two types. The prior beliefs held by Player 1 about Player 2's types ($t_2$ = 1 or 2) are given by the probabilities 0.6 and 0.4, respectively.

All the above models simply describe different situations as games. We have not made any attempt to analyze the games to find what rational players should or would do while playing these games. The next section is devoted to such analyses.

## Solutions

We are now ready to analyze our games. We look for a "solution" (a specific outcome) of a game. It may not be a prescription in the sense that the solution does not really tell us how to play a given game. We, the theorists, are rather trying to understand how (rational and intelligent) players think and play games.

The strategy profile associated with the solution-outcome of a game presents an "equilibrium" behavior. The advancement of game theory is indeed demonstrated in the power of theoretical solutions to explain observed behavior and in many laboratory-based experiments where players follow equilibrium behavior. The critics, however, often claim that theoretical solutions fail to offer any prediction in some experiments and thereby question these equilibrium concepts. For a balanced and advanced discussion, see Camerer (2003). For an alternative approach to justify theoretical notions, one may consider "revealed preference in games" (Ray & Zhou, 2001).

To find an acceptable solution, we must first consider the concept of dominance. A strategy is dominated if it generates lower payoffs than some other, dominating strategy of that player, whatever the opponents' strategies. For example, in the $2 \times 2$ game in Figure 17.12 (only Player 1's payoffs are shown), the strategy *B* for Player 1 is (strictly) dominated by the dominant strategy *T* as *T* generates (strictly) higher payoffs.

A dominant strategy equilibrium is a strategy profile where each player is playing a dominant strategy. For example, in the prisoner's dilemma game discussed earlier (in Figure 17.2), (Defect, Defect) is the dominant strategy equilibrium as Defect is the dominant strategy for both players.

Unlike the prisoner's dilemma, most games do not possess any dominant strategy equilibrium. Some games, however, can be solved by a process of "iterative elimination of dominated strategies." As the players do (should) not play dominated strategies, we eliminate them from the game one by one and thus reach a solution or an equilibrium of the game.

For example, in the two-person game in Figure 17.13, the strategy $R$ for Player 2 is dominated by the strategy $M$ and thus can be deleted from the game. Once $R$ is deleted, in the rest of the game, the strategy $T$ for Player 1 is dominated by the strategy $B$ and hence should be deleted. In the rest of the game, clearly, the strategy $M$ is dominated for Player 2 and hence can be deleted. Thus, the outcome $(B, L)$ turns out to be the solution of this game based on the process of iterative elimination of dominated strategies.

It should be pointed out here that the order of elimination may matter and may produce different solutions for different orders used, if the process involves weak dominance (where the payoffs in the dominated strategies are weakly worse).

One may also consider domination by a randomized strategy. For example, in the game in Figure 17.14 (only Player 1's payoffs are shown), the strategy $M$ for Player 1 can be dominated if Player 1 is allowed to choose the strategies $T$ and $B$ with equal probabilities. In this game, this random choice of Player 1 will generate an expected payoff of 0.5, which is higher than that from the strategy $M$ for him or her. Thus, $M$ is dominated by the strategy $(½T, ½B)$.

Strategies such as $(½T, ½B)$ are called "mixed" strategies. Mixed strategies are probability distributions over the pure strategies. These are randomly chosen by a player according to the probability distribution. They can be also interpreted as beliefs held by a player over the other player's strategies.

Allowing for mixed strategies, a (normal-form) game $G = < N, S_1, \ldots, S_n, U_1, \ldots U_n >$ can now be extended to $< N, \Sigma_1, \ldots, \Sigma_n, u_1, \ldots, u_n >$ in which $\Sigma_i$ denotes the set of all mixed strategies of player $i$, where $\sigma_i \in \Sigma_i$ is a typical mixed strategy and $u_i$ is the expected utility of player $i$. To find the expected payoff from a profile of mixed strategies, note that the mixed strategies are chosen "independently," and hence the joint probability of a particular pure profile $s = (s_1, \ldots, s_n) \in S$ is simply the product of the probabilities $\sigma_i(s_i)$. Hence, $u_i(\sigma_1, \ldots, \sigma_n) = \Sigma_{s \in S} [\prod_i \sigma_i(s_i)] U_i(s)$.

For extensive-form games, randomization can be viewed in two ways. In such games, as mentioned earlier, a strategy is a complete plan of actions indicating what a player does in each of his or her information sets. A mixed strategy thus is a distribution over these complete plans. For example, in the umbrella game, a mixed strategy for Player 1 is a probability distribution (say, ¼ each) over the four pure strategies *UU*, *UN*, *NU*, and *NN*. Clearly, mixed strategies defined this way are difficult to understand and

(a)

| | *L* | *R* |
|---|---|---|
| *T* | 1, 2 | 0, 1 |
| *B* | 0, 4 | 1, 3 |

(b)

| | *L* | *R* |
|---|---|---|
| *T* | 1, 3 | 0, 4 |
| *B* | 0, 1 | 1, 2 |

**Figure 17.11** A Bayesian Game

*Notes:* (a) $t_2 = 1$ (with probability 0.6). (b) $t_2 = 2$ (with probability 0.4).

| | *L* | *R* |
|---|---|---|
| *T* | 1, ... | 5, ... |
| *B* | 0, ... | 4, ... |

**Figure 17.12** A Game With a Dominant Strategy for Player 1

| | *L* | *M* | *R* |
|---|---|---|---|
| *T* | 4, 3 | 2, 7 | 0, 4 |
| *B* | 5, 5 | 5, –1 | –4, –2 |

**Figure 17.13** A Game Solvable by Iterative Elimination of Dominated Strategies

| | *L* | *R* |
|---|---|---|
| *T* | 2, ... | –1, ... |
| *M* | 0, ... | 0, ... |
| *B* | –1, ... | 2, ... |

**Figure 17.14** A Game Where a Randomized Strategy of Player 1 Is Dominant

interpret. Instead, one may work with another way of randomizing called behavioral strategies, which allow mixing over actions within each of the information sets for each player. For example, in the umbrella game, Player 1 may play a behavioral strategy that randomizes (say, ½–½) over the actions *U* and *N* when it rains and randomizes (say, ¼–¾) when it is sunny. Behavioral strategies clearly are easy to interpret and handle. Moreover, it can be technically proved that mixed and behavioral strategies are mathematically "equivalent."

Not many games can be solved by eliminating dominated strategies. Hence, we need other concepts to find an acceptable solution. The most popular equilibrium solution concept in game theory is the Nash equilibrium, named after John Nash, who introduced this notion (Nash, 1950b). Before we define this equilibrium notion for a (normal-form) game, we first focus on the concept of "best response" of any player against his or her opponents' strategies.

For example, in the game of chicken (as in Figure 17.5), the best response for a player can be easily demonstrated. The best response for a player in this game is the strategy *P* if the opponent plays the strategy *A* and vice versa.

In a Nash equilibrium strategy profile, every player plays the best response against the other players' strategies specified in the profile. The interpretation is that in equilibrium, everybody is playing the best possible strategy, given that the others are also playing their best strategies. Assuming everyone else's strategy, no player can be (strictly) better off by deviating unilaterally. Clearly, a solution achieved through the elimination of dominated strategies is also a Nash equilibrium.

Formally, $s^*$ (a strategy profile) is a Nash equilibrium if for all $i$, $u_i(s^*) \geq u_i(s_i, s^*_{-i})$, for all other $s_i$. For example, in chicken, $(A, P)$ is a Nash equilibrium.

For extensive-form games, where players move sequentially, one may use Nash's notion of equilibrium by treating players' strategies as complete plans of actions before play begins. Also, one can find the Nash equilibrium in behavioral strategies for extensive-form games by checking the equilibrium condition for each player.

For Bayesian games, a Bayesian Nash equilibrium is a strategy profile (one strategy for each type of each player) such that each type is maximizing (expected) payoff given others' strategies. For example, in the game in Figure 17.11, the profile $[(T; (L, R)]$ is a Bayesian Nash equilibrium. Indeed, in this game, the strategies *L* and *R*, respectively, are dominant for the two types of Player 2.

Unfortunately, there may be games for which there are no Nash equilibria in pure strategies. For example, in the matching pennies game (as in Figure 17.1), none of the four outcomes is a pure Nash equilibrium.

One may define, similarly, Nash equilibrium in mixed strategies. $\sigma^*$ is a Nash equilibrium if, for all $i$, $\sigma^*_i$ is the best response against $\sigma^*_{-i}$. John Nash has proved the theorem that a Nash equilibrium in mixed strategies always exists. Nash equilibrium in pure strategies exists under some conditions (the proof of these theorems uses a mathematical "fixed-point theorem").

It can be mathematically proved that for any mixed strategy $\sigma_i$ to be the best response for any player $i$, it must be the case that all the pure strategies in its support (on which player $i$ is mixing) are best responses as well. Thus, the player has to be indifferent in equilibrium (otherwise he or she would not mix). Using this technical result, in a $2 \times 2$ game, we can get two indifference equations, solving which we can find the Nash equilibrium in mixed strategies.

For games with a continuum of strategies, one can find the best response of a player using the first-order conditions. The Cournot (Nash) equilibrium for an oligopoly market is an example of this.

## Refinements

The Nash logic ignores the sequential and the informational structure in extensive-form games and, therefore, sometimes may not be justifiable. Consider the entry deterrence game as in Figure 17.15, in which there are two firms (players); one is a possible entrant (Player 1) to an industry in which there is an incumbent (Player 2). The potential entrant moves first and has two choices, to enter (*E*) and to stay out (*O*). If it does not enter, the game ends, while if it enters, the incumbent then has two options: to accommodate (*A*) and to fight (*F*).

The game can be analyzed in its normal form as described by the table in Figure 17.16.

The above game has two pure Nash equilibria, $(O, F)$ and $(E, A)$. Indeed, in the equilibrium profile $(O, F)$, the

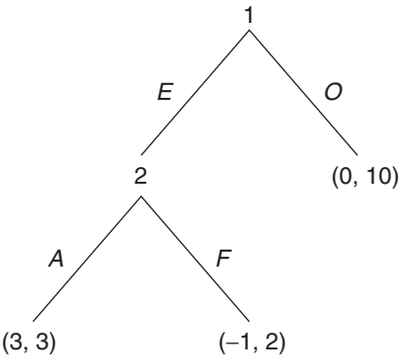


**Figure 17.15** An Entry Game

| | *A* | *F* |
|---|---|---|
| *E* | 3, 3 | –1, 2 |
| *O* | 0, 10 | 0, 10 |

**Figure 17.16** Normal Form of the Entry Game

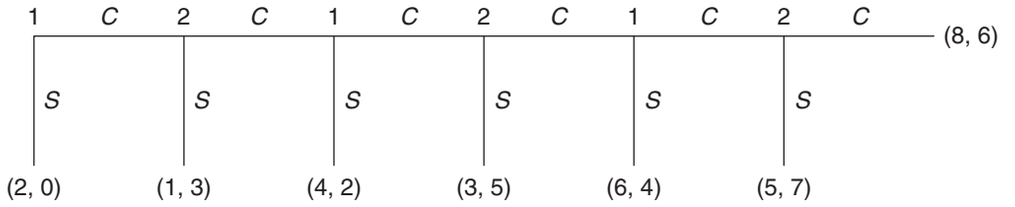


**Figure 17.17** A Centipede Game

entrant is playing the best response *O*, as it believes that the incumbent will choose *F*. However, this belief is based on an "incredible threat." *F* is not optimal in the unreached "subgame" (intuitively, the "game" below a nonterminal node) in which the incumbent has to make a choice (the incumbent did not make any choice here as the game ended when the entrant chose *O*).

A subgame perfect equilibrium, as defined by Selten (1965), is a strategy profile that induces a Nash equilibrium in every "subgame" of the original game, even if it is off the equilibrium path. Clearly, subgame perfect equilibrium refines the set of Nash equilibria. In the above example, (*E*, *A*) is a subgame perfect equilibrium, while (*O*, *F*) is not.

Any finite extensive-form game with perfect information (no information set in the tree) can be solved using "backward induction." This induction process finds the optimal actions of the players in the "last" subgame first and then, given these actions, works backwards to the beginning to find the subgame perfect equilibrium of the game.

Consider the two-player "centipede" game as in Figure 17.17, in which each player sequentially chooses either to continue (*C*) or to stop (*S*). Backward induction implies that the unique subgame perfect equilibrium is to play *S* for both players in every stage. Thus, the game should be stopped in the very first stage. This outcome has been the center of attention of a long debate (for a recent update on this, see Palacios-Huerta & Volij, 2009).

For games with imperfect information, just the notion of subgame perfection may not be enough to justify a solution. The concepts of perfect Bayesian equilibrium and sequential equilibrium were introduced to incorporate a set of beliefs over all information sets. These solutions consist of strategy profiles and beliefs such that each player is playing the best strategy given the beliefs (sequential rationality), and the beliefs are consistent, both on and off the equilibrium path (Kreps & Wilson, 1982).

## Correlation

In noncooperative games, how one can generate strictly higher (expected) payoffs than the Nash equilibrium payoffs is an important and interesting question. One answer lies in *correlation* among players.

The concept of correlation in games can best be associated with the notion of *mediation*, as often seen in many real-life political scenarios, where decision makers go to an independent and nonstrategic mediator and choose their actions based on the suggestions of the mediator. A (direct) correlation device can be identified as a mediator who can help the players to play a given game. The mediator (correlation device) randomly picks an outcome (a strategy profile) of the game, according to a probability distribution, and privately recommends to the players to play the respective part of the outcome.

A (direct) *correlated equilibrium* (Aumann, 1974, 1987) is a mediator whose recommendations the players find optimal to follow obediently—that is, following the mediator's recommendations constitutes a Nash equilibrium of the extended game. Correlation typically improves upon Nash equilibria of the original game; notable exceptions are zero-sum games and a class called "strategically zero-sum" games identified by Moulin and Vial (1978). The notion of correlation can be suitably extended to finite games with incomplete information (Forges, 1993, 2006).

For example, consider the game of chicken as in Figure 17.5 and the correlation device as in Figure 17.18.

The correlation device is nothing but a probability distribution, *P*, over the four pure outcomes of the game. The device first selects a "cell" and sends private messages, and the players play the game. In the extended game, each player has four strategies out of which "follow obediently" is a strategy. The device is called a correlated equilibrium if the obedient strategy is a Nash equilibrium in the extended game.

Formally, a device *P* is called a correlated equilibrium if, for all $i$, for all $s_i \in S_i$, $\Sigma_{s-i} P(s_{-i}, s_i)\, u_i(s_i, s_{-i}) \geq \Sigma_{s-i} P(s_{-i}, s_i)\, u_i(t_i, s_{-i})$, for all other $t_i \in S_i$ (that is, $s_i$ is the best response for player $i$ against the opponents' strategies of following their recommendations).

| | *A* | *P* |
|---|---|---|
| *A* | 0 | 1/3 |
| *P* | 1/3 | 1/3 |

**Figure 17.18** A Correlation Device for Chicken in Figure 17.5

Note that there may be multiplicity of equilibria; that is, obedience is just one Nash equilibrium, but other equilibria may exist.

One may also extend mediation to Bayesian games that are described by players, types, and strategy sets. A mediated mechanism, denoted by $\mu$, is simply a map from $T = \prod_i T_i$ where $T_1, \ldots, T_n$ are respectively the set of types for each player to the probability distributions over $S$, denoted by $\Delta(S)$. In this extended game, the players report types (or messages) to the mechanism, get recommendations (or messages) from the mechanism, and then play the original game.

A strategy of a player in the extended game, extended by the mechanism, is to report a type and then play a strategy in the game. A mechanism $\mu$ is called a "communication equilibrium" if honesty and obedience is a Nash equilibrium in the extended game.

Policy making can be associated with designing a mechanism that is incentive compatible (to satisfy the equilibrium condition above) and individually rational (implying that the payoffs from the mechanism are no worse than the outside option, normalized to zero, that is, $u_i(\mu \mid t_i) > 0$, for all $i$, for all $t_i \in T_i$).

## Repetition

The motivation for studying repeated games comes from the prisoner's dilemma game as discussed earlier (Figure 17.2).

The question is how to achieve the socially optimal outcome (Cooperate, Cooperate) as an equilibrium outcome. The answer lies in repetition of the same game again and again. If we interact repeatedly, we may introduce a social norm that incorporates nice behavior and punishment.

The research in repeated games has indeed achieved the desirable outcome as an equilibrium and found the structure of the strategies (behavior) to support this. Also, the different forms of behavior can be analyzed as finite machines (automata).

Let $G = < N, S_1, \ldots, S_n, U_1, \ldots, U_n >$ be a basic normal-form game that can be repeated, possibly, infinitely many times. This is unlike the bargaining game discussed earlier where the game may potentially go on for infinitely many periods. If $G$ has a unique Nash equilibrium, then the subgame perfect equilibrium of the finitely repeated version of this game is the Nash equilibrium of $G$. For example, for a finitely repeated prisoner's dilemma, the only equilibrium is (Defect, Defect) in every period.

In an infinitely repeated game, a strategy of player $i$ is a function that assigns a strategy in $S_i$ to every sequence of outcomes (history), and the preferences are defined over (infinite) sequences of outcomes (numbers).

The result (folk theorem) suggests that every feasible individually rational payoff can be supported as a Nash equilibrium outcome. The proof of the theorem uses the notion of a "trigger" strategy that cooperates until the opponent defects and then punishes forever.

One may analyze strategies for any infinitely repeated game using finite machines or automata (Rubinstein, 1986). The interpretation is that a player implements a strategy using a machine. For example, the trigger (also called the "grim") strategy can be described as a machine that has two states. In the first state of the machine, the player cooperates. The machine moves to the second "absorbing" state only when the opponent defects and defects forever.

## Cooperation

Cooperative game theory studies coalitions (groups) and how they share the benefits, assuming that the worth of a coalition can be shared among the members and that the utility is "transferable."

These models are known as games in coalitional form. The literature does not ask how the coalitions form or which coalition is more likely to form. We assume all possible coalitions are available. The ingredients of this model are coalitions and their values.

Let $N$ be the set of players as in noncooperative games; thus, the possible coalitions, including singletons (individual players), are $(2^N - 1)$. The value of any coalition $C$, given by the characteristic function $v$, is a map from $(2^N - 1)$ to the set of real numbers, with $v(C)$ denoting the value of coalition $C$. A game in coalitional form thus has a complete description given by $(N, v)$.

For example, the three-person majority game is described by $N = \{1, 2, 3\}$; $v(\{i\}) = 0$; $v(C) = 1$ for any other $C$ (implying that the majority wins).

Similarly, consider a situation with a seller (Player 1) who has an object to sell to two possible buyers (Players 2 and 3) with valuations 90 and 100, respectively. The game can be described by $v(\{i\}) = v(\{2, 3\}) = 0$; $v(\{1, 2\}) = 90$; $v(\{1, 3\}) = 100$; $v(\{1, 2, 3\}) = 100$.

To find a solution for such games, we look for allocations (also called imputations), $x$, that are vectors $(x_1, \ldots, x_n)$ such that $\sum_i x_i = v(N)$ and $x_i \geq v(\{i\})$. Thus, allocations divide the whole benefit such that the division is individually rational.

We say that an allocation $x$ dominates another allocation $y$ through coalition $C$ if $x_i > y_i$ for all $i \in C$ and $\sum_{i \in C} x_i \leq v(C)$ (members in the coalition $C$ are strictly better off). A coalition $C$ blocks $y$ if there exists $x$ such that $x$ dominates $y$ through $C$. For example, in the three-person majority game, (½, ½, 0) is blocked by the coalition (2, 3) with an allocation (0, 2/3, 1/3).

A desirable allocation is one that is acceptable to all coalitions, that is, not blocked by any coalition. Hence, our first "solution" for these games is indeed an allocation that is not blocked. There may be many such unblocked allocations for a game. Thus, the solution is the set of undominated allocations (the allocations that no coalition can block). This set is called the core.

It can be shown that an allocation $x \in$ core if and only if $\sum_i x_i = v(N)$ and $\sum_i x_i \geq v(C)$ for all $C$. Thus, from the

above characterizing inequalities, the core is a closed and convex set. The core for some games may be empty. For example, the core is empty for the three-person majority game. The core of the seller and buyers game is given by $\{t, 0, 100 - t \mid 90 \leq t \leq 100\}$.

Another solution concept for these games is the ($vN - M$) *stable set*, named after Von Neumann and Morgenstern. This concept is based on desirable properties of a set of allocations.

A set of allocations, $V$, is called *internally stable* if for any two allocations $x, y \in V$, $x$ does not dominate $y$; that is, no two allocations in the set dominate each other. A set of allocations, $V$, is called *externally stable* if for any allocation $z \notin V$, there exists an allocation $x \in V$ such that $x$ dominates $z$ (through some coalition $C$); that is, any allocation outside the set is dominated by something in the set.

A set $V$ of allocations is called the ($vN - M$) *stable set* if the set demonstrates the two "stability" properties above—that is, if it is both internally and externally stable.

It is easy to show that for the three-person majority game, {(½, ½, 0), (½, 0, ½), (0, ½, ½)} is a stable set. Stable sets, however, may not exist for some games, as proved in a 10-person example by Lucas (1969).

## Conclusion

In this chapter, we discussed all the models and some solutions in game theory. Needless to add, there are indeed many other concepts that we have not discussed. Readers are encouraged to consult the list of textbooks in game theory below for further references. Also, there are many specific advanced topics (such as evolution and learning) and applied topics (such as bargaining, auctions) where the models and solution concepts have been used and applied. For example, alternating-offers bargaining between two parties (players) can be modeled as a noncooperative game (Rubinstein, 1982) in an extensive form and solved using the notion of subgame perfection. Sealed-bid auctions among several bidders (players) with each player $i$ having a private valuation ($v_i$) of the object can be described as a Bayesian game and solved using Bayesian-Nash equilibrium. The list below also includes some advanced texts on these specialized topics.

*Author's Note:* The author gratefully acknowledges many extremely helpful comments and constructive suggestions received from an anonymous referee, Ralph Bailey, Chirantan Ganguly, and Sonali Sen Gupta and particularly the support provided by the editor and her team.

## References and Further Readings

The following is a list of texts that readers may find useful. The first section lists general textbooks while the rest are on specific topics. The final section lists sources cited in the chapter.

### General

Bierman, H. S., & Fernandez, L. (1997) *Game theory with economic applications* (2nd ed.). Reading, MA: Addison Wesley.

Binmore, K. (1991). *Fun and games: A text on game theory.* Lexington, MA: D. C. Heath.

Binmore, K. (2007). *Playing for real: A text on game theory.* Oxford, UK: Oxford University Press.

Carmichael, F. (2005). *A guide to game theory.* Upper Saddle River, NJ: Prentice Hall.

Dixit, A., & Skeath, S. (2004). *Games of strategy.* New York: W. W. Norton.

Dutta, P. K. (1999). *Strategies and games: Theory and practice.* Cambridge: MIT Press.

Fudenberg, D., & Tirole, J. (1991). *Game theory.* Cambridge: MIT Press.

Gardner, R. (1995). *Games for business and economics.* New York: John Wiley.

Gibbons, R. (1992). *A primer in game theory.* Englewood Cliffs, NJ: Prentice Hall.

Holt, C. A. (2007). *Markets, games and strategic behavior.* Boston: Pearson.

McCain, R. A. (2004). *Game theory: A non-technical introduction to the analysis of strategy.* Mason, OH: Thomson/South-Western.

Montet, C., & Serra, D. (2003). *Game theory and economics.* New York: Palgrave.

Myerson, R. B. (1991). *Game theory: Analysis of conflict.* Cambridge, MA: Harvard University Press.

Osborne, M. J. (2004). *An introduction to game theory.* Oxford, UK: Oxford University Press.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory.* Cambridge: MIT Press.

Rasmusen, E. (2001). *Games and information.* London: Blackwell.

Romp, G. (1997). *Game theory: Introduction and applications.* Oxford, UK: Oxford University Press.

Vega-Redondo, F. (2003). *Economics and the theory of games.* Cambridge, UK: Cambridge University Press.

Watson, J. (2002). Strategy: *An introduction to game theory.* New York: W. W. Norton.

### Auctions

Klemperer, P. (2004). *Auctions: Theory and practice.* Princeton, NJ: Princeton University Press.

Krishna, V. (2002). *Auction theory.* New York: Academic Press.

Menezes, F. M., & Monteiro, P. K. (2005). *An introduction to auction theory.* Oxford, UK: Oxford University Press.

Milgrom, P. R. (2004). *Putting auction theory to work.* Cambridge, UK: Cambridge University Press.

### Bargaining

Osborne, M. J., & Rubinstein, A. (1990). *Bargaining and markets.* New York: Academic Press.

Muthoo, A. (1999). *Bargaining theory with applications.* Cambridge, UK: Cambridge University Press.

### Evolution and Learning

Cressman, R. (2003). *Evolutionary dynamics and extensive form games.* Cambridge: MIT Press.

Fudenberg, D., & Levine, D. (1998). *Learning in games.* Cambridge: MIT Press.

Samuelson, L. (1998). *Evolutionary games and equilibrium selection.* Cambridge: MIT Press.

Vega-Redondo, F. (1996). *Evolution, games and economic behavior.* Oxford, UK: Oxford University Press.

Weibull, J. W. (1995). *Evolutionary game theory.* Cambridge: MIT Press.

Young, H. P. (1998). *Individual strategy and social structure.* Princeton, NJ: Princeton University Press.

Young, H. P. (2004). *Strategic learning and its limits.* Oxford, UK: Oxford University Press.

## References

Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics, 1,* 67–96.

Aumann, R. J. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica, 55,* 1–18.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction.* Princeton, NJ: Princeton University Press.

Forges, F. (1993). Five legitimate definitions of correlated equilibrium in games with incomplete information. *Theory and Decision, 35,* 277–310.

Forges, F. (2006). Correlated equilibrium in games with incomplete information revisited. *Theory and Decision, 61,* 329–344.

Kreps, D. M., & Wilson, R. B. (1982). Sequential equilibria. *Econometrica, 50,* 863–894.

Lucas, W. F. (1969). A game in partition function form with no solution. *Siam Journal on Applied Mathematics, 16,* 582–585.

Moulin, H., & Vial, J. P. (1978). Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory, 7,* 201–221.

Nash, J. (1950a). The bargaining problem. *Econometrica, 18,* 155–162.

Nash, J. (1950b). Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences, USA, 36,* 48–49.

Palacios-Huerta, I., & Volij, O. (2009). Field centipedes. *American Economic Review, 99,* 1619–1635.

Ray, I., & Zhou, L. (2001). Game theory via revealed preferences. *Games and Economic Behavior, 37,* 415–424.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica, 50,* 97–110.

Rubinstein, A. (1986). Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory, 39,* 83–96.

Selten, R. (1965). Spieltheoretische behandlung eines oligopolmodells mit nachfragetragheit [A game-theoretic approach to oligopoly models with inert demand]. *Zeitschrift fur die Gesamte Staatswissenschaft, 121,* 301–324, 667–689.

Von Neumann, J., & Morgenstern, O. (1944). *The theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

# 18

# Economics of Strategy

Quan Wen
*Vanderbilt University*

Economic and other rational decision making involves choosing one of the available actions for the environment that the decision maker is facing, given his or her objectives. Rational decision making and the subsequent choice of action depend on these two primary factors. Economics of strategy is concerned with how a rational decision maker's behavior in choosing an action can be explained by logical and strategic reasoning. A simple decision problem can be viewed as the simplest strategic problem, when there is only one agent for a given environment. Many economic problems, however, involve multiple decision makers. Our primary goal in economics of strategy is to provide a systematic analysis of how each decision maker behaves and how the strategic logics of the decision makers interact.

The primary tool researchers use in strategic analysis is game theory. Game theory provides us with a systematic and logical method for studying decision-making problems that involve multiple decision makers. Game theory has been applied in many areas of economic studies, such as industrial organization and strategic international trade. Application of game theory has different implications, depending on the specific economic environment. To facilitate our understanding of how rational and strategic decisions are made, we focus on firms' strategic behavior in settings with different market structures. We review necessary and commonly used concepts and modeling techniques of game theory and apply them in analyzing firms' strategic behavior.

## Competitive Market and Competitive Firms

We usually do not think that an individual firm behaves strategically in a competitive market; after all, a competitive firm simply chooses a feasible action to maximize its profit. However, a competitive firm's decision problem can be viewed as a strategic problem with only one agent. In this section, we review how a competitive firm behaves in reaction to market conditions and how this reaction can be considered a simple form of strategic behavior. Some of the basic arguments we discuss in this section will help us understand more complicated strategic interactions under different market structures.

Generally speaking, a competitive firm does not have market power to influence the market price. One interpretation of this simplified assumption is that there are many similar firms in the market, and each individual firm is too small to have any significant impact on the market outcome. When making a decision, such as how much to produce in the upcoming production period, a competitive firm, however, still needs to evaluate the economic environments it will face. The main factors that dictate how a competitive firm behaves are the prices of its inputs and outputs. For example, when the price of an input increases, firms decrease output. On the other hand, when the economy grows and the price of the output increases, each individual producer will increase its output as a response to a higher price. When a firm decides how much to produce, it will weigh the benefit and cost of its primary activity. To a competitive firm, the benefit of producing one more unit of its output is its marginal revenue, which is given by the market price, while its cost for producing one more unit of its output is its marginal cost. In order to achieve the maximal profit, a competitive firm will choose to produce in such a way that its marginal cost of production is equal to its marginal revenue—that is, the market price. If a firm's marginal cost is not equal to the market price of its output, the firm could produce at a different output level to achieve a

higher profit. For example, if the market price were higher than its current marginal cost, the firm would be able to increase its profit by increasing its production. Then, how much a firm will produce is determined by the market price and the firm's reaction to the market price of its products. Alternatively, if we consider that the feasible action of an individual firm is to choose its level of inputs, then the same principle applies here but for different types of activities; the marginal benefit of a production factor is the market value of the marginal product of the factor, and the marginal cost of the factor is simply its market price.

No matter whether we consider choosing outputs or choosing inputs as the primary decision for an individual firm, the firm needs to adjust its decision for the economic environment it will be in. This adjustment can be viewed as a very simple form of strategic behavior—namely, actions in response to the environment. In other words, the behavior of a competitive firm is a plan or a course of actions conditional to economic environments and conditions; it is a strategy of the firm in this degenerated strategic environment with only one agent. We will see how this line of argument applies in more general forms of strategic setting when multiple agents interact, such as in an oligopoly market.

Another type of decision problem a firm may encounter in a competitive market is to decide whether to enter the market. The crucial factor that dictates a firm's decision to enter a market is its projected profitability. If a firm's projected profit from entering a market is more than any cost associated with entering the market, the firm will enter the market, which will lower the profitability of the other firms in the market.

## Monopoly and Monopolistic Behavior

The other extreme market structure is monopoly, in which there is only one firm that accounts for all production and sales in the entire market. In this situation, the monopolistic firm has full market power to determine how much to produce and how much to charge for its product. This is another decision problem because there is only one economic agent, namely, the monopolistic firm. Like a competitive firm, the monopolistic firm will balance the cost and benefit of its production activities in order to achieve a higher profit. In an ordinary monopoly market, the firm can choose how much to produce and let the market demand determine the price of its product. Alternatively, the firm may choose how much to charge for its product, and the market demand will determine how much the firm will be able to sell. Both models will lead to the same result for an ordinary monopoly. Researchers commonly work with the model in which the firm chooses how much to produce.

To maximize its economic profit, the firm needs to produce in such a way that its marginal cost of production is equal to its marginal revenue. The marginal cost of production is determined by the production technology of the firm, as well as the market structures of its inputs. The firm's marginal cost generally varies with respect to how much the firm produces, whether it is a competitive firm or a monopolistic firm. However, unlike in a competitive market, the marginal revenue of the firm in a monopoly varies with its output level. Producing more output has two opposite effects on the firm's profit. On one hand, additional production will generate more revenue for the firm. On the other hand, additional production will lower the price of its product, and hence, it has a negative impact on revenue. Even though the firm's marginal revenue varies with its output level, the monopolist should choose the output level at which the marginal revenue and the marginal cost are the same; this is a general principle for profit maximization. For different market structures, marginal cost and marginal revenue are determined differently, although the general principle for profit maximization is the same.

For a nonordinary monopoly, price discrimination is an important subject to investigate. There are mainly two sets of characteristics, quantity and identity, on which the monopolist may be able to discriminate in setting up different prices. In first-degree price discrimination, the monopolist charges different per-unit prices for different consumers or different groups of consumers, but this per-unit price is independent of how much the consumer purchases. Under first-degree price discrimination, we can separate the firm's profit maximization problem into different problems for different consumers or different groups of consumers, and in each separated market, the problem is equivalent to an ordinary monopoly problem with different market demands. In second-degree price discrimination, the monopolist is unable to set different prices for different consumers or different groups of consumers but can set different prices for different quantities any consumer buys. Volume discounts represent a case of second-degree price discrimination. If the monopolist can set different prices for different consumers and for different quantities, the practice is referred as to third-degree price discrimination or perfect price discrimination. With any form of price discrimination, the firm will have a higher profit than its counterpart in an ordinary monopoly.

## Static Oligopoly and Strategic Interaction

Market structure that falls between perfect competition and monopoly is known as oligopoly, in which there are a number of firms that serve the market. Unlike in perfect competition, each individual firm in an oligopolistic market influences market outcome through its production activities, but unlike in monopoly, the market outcome also depends on what the other firms in the market do. The outcome of such an oligopolistic competition depends on how all the firms compete jointly in the

market. The problem faced by each individual firm is no longer a simple decision problem, but rather a situation of strategic interaction. Recall that in a decision problem, the decision maker chooses one of the feasible actions given the environment. Although what is optimal depends on the economic environment or economic condition, the economic environment does not respond to what feasible action the decision maker chooses. In a monopoly, for example, no matter how much the firm produces or how much it charges for its product, the economic environment, which is summarized by the demand function, does not change with respect to the firm's decision. In oligopoly, however, when an individual firm decides its output or price, it is necessary to take into account how the other firms will respond to its decision, because the other firms will choose their competitive strategies based on the environments they are facing, which, in turn, depend on what competing strategy the firm under consideration will take. Because of this strategic interaction among the firms, the tool we use to analyze oligopoly problems is game theory. Before we come back to this basic static oligopoly problem, we first review some of the basic notions and concepts of game theory.

## Strategic Form Game and Nash Equilibrium

Strategic form representation, also known as the normal form representation, of a game consists of three basic and necessary ingredients: who plays the game, what each individual in a game can do, and what each individual cares about. For simplicity, we often refer to an individual in a game as a player. The interpretation of a player is rather broad. For example, a player can be an individual firm in an oligopolistic market or a country in an international trade problem. The second component in the strategic form representation of a game is the feasible actions of the players. We often interpret a strategic form game as a static game, in which all players choose their actions or strategies simultaneously. When choosing a strategy, a player does not have any information on which strategies the other players play. In this type of game model, strategies are the same as actions; a player simply chooses an action. When we discuss how to model a dynamic game, we detail the differences between strategies and actions. Given one strategy from every player, an outcome of the game emerges. Players have preferences over the set of all possible outcomes of the game, and this is the last component of a strategic game.

Because a player cares about the outcome of the game and the outcome is determined jointly by the strategies of all the players, a player cannot guarantee that a particular outcome of the game will prevail; this player must take into account the strategies used by the others. If we fix the strategies adopted by the other players, then the problem faced by the player under consideration becomes a simple decision problem, and we can derive the best action for this player to choose. Of course, what is considered optimal for this player depends on the strategies of all the other players, and this dependence is described as the best response or reaction function. This best response is a crucial step to formulating the equilibrium concept in a strategic form game.

A player's best response is stable in the sense that the player does not have any incentive to play a strategy that is not one of the best responses to the strategies chosen by the other players. Nash equilibrium is one of the widely used solution concepts in studying strategic form games. A strategy profile—that is, a list of players' strategies, one from each player, is called a Nash equilibrium if any player's strategy in the profile is a best response to the other players' strategies in the profile. There is one way to interpret this equilibrium concept: When every player decides which strategy to play, the player first forms an expectation of what the other players will play, then chooses one of the best responses to the expectation of the other players' strategies. Of course, in order for a player's expectation of the other players' strategies to be consistent, this player's expectation should be correct in the equilibrium. These two criteria in a Nash equilibrium determine what outcomes should be considered equilibrium outcomes and which should not. If an outcome, or the strategy profile that induces the prescribed outcome, is not a Nash equilibrium outcome, then there must exist at least one player who has incentive to deviate to one of the other feasible strategies, given what the other players would play in the strategy profile. Such an outcome is not stable, and hence, it should not be considered an equilibrium outcome.

## The Cournot Model

In the rest of this section, we review two classic static oligopoly models—namely, Cournot competition and Bertrand competition. In the Cournot model, firms compete in their outputs, firms simultaneously choose how much to produce, and then the market price is determined by the inverse market demand function. With or without product differentiation, this Cournot model can be considered a standard game in strategic form, in which the firms are players, their feasible output levels are the sets of strategies, and the firms' objectives are to maximize profits.

As in a generic strategic form game, in order to identify possible equilibrium outcomes, we must figure out each firm's best response in this Cournot model. When an oligopolistic firm decides how much to produce, it needs to have an expectation of how much the other firms will produce. Given this belief or expectation, the problem faced by an individual firm becomes a relatively simple decision problem. There is a unique correspondence between the market price and the output of the firm under consideration. The more the firm produces, the lower the market price for its product. Each oligopolistic firm faces a dilemma similar to that of a monopolist, and in choosing

how much to produce, it must balance the costs and benefits. When solving this firm's profit maximization problem, we treat the outputs of the other firms as a given. Then, the optimal output level of the oligopolist necessarily depends on how much the others produce. The relation between the optimal output of a firm and the output levels of the other firms provides us with the best response of this firm under consideration. Repeating the same process for every firm, we can identify the oligopolist's best response to the outputs of the other firms. An outcome that satisfies all the best responses of all firms is a Nash equilibrium outcome in the Cournot model, which is also commonly known as Cournot equilibrium.

We characterize Cournot equilibrium by taking account of strategic interaction among the firms. Because every firm's strategy in the equilibrium is a best response to the other firms' strategies in the profile, every firm's strategy in Cournot equilibrium is rationalizable. A strategy is rationalizable if it is a best response to the other strategies that are also rationalizable, and are the best responses of some other rationalizable strategies. Generally speaking, rationalizable strategies include all equilibrium strategies.

Another idea to highlight the strategic interaction among the firms is the dynamic adjustment process to reach equilibrium. For any strategy profile, if it is not an equilibrium, some firms may want to revise their strategies in the profile. After this revision of strategy, some other firms may want to revise their strategies, or some of the firms that revised their strategies may want to revise again. This adjustment process may or may not converge. When it does, this progress will converge into a stable equilibrium. In a nonstable equilibrium, if one firm produces a slightly different amount of output than the one given in the strategy profile, this adjustment process may diverge or converge into a different strategy profile. These arguments help us understand how firms interact when competing in a Cournot competition.

## Collusion and Prisoners' Dilemma

Taking the basic setup of the Cournot model, we can examine collusive outcomes. In a departure from the concept of Nash equilibrium, in which every firm maximizes its own profile, collusion happens when the firms form a cartel and coordinate their production strategies to maximize their aggregate profits. To maximize firms' joint profits, each firm will set its marginal cost equal to the marginal revenue of all the firms, which is the rate of change of firms' joint profit with respect to the output of an individual firm. Because the marginal revenue of the cartel is less than the marginal revenue of an individual firm, every firm may increase its profit by producing more, if all the other firms produce their quotas in the cartel. In other words, a collusive outcome is not an equilibrium outcome in this static setting because every firm has an incentive to deviate from a cartel agreement to produce more than what is supposed to be produced in the cartel.

The static Cournot model closely resembles a classic strategic game known as the prisoners' dilemma. In a prisoners' dilemma, two suspects are interrogated separately and simultaneously by the police for the crime they committed. To gather sufficient evidence for a successful conviction, the police offer leniency to the suspect who provides information, if and only if the other suspect refuses to do so. Each suspect can take one of two possible actions or strategies, either to collude by not providing any information to the police or to defect to the police. Regardless of whether the other suspect chooses to collude or defect, a suspect will always receive a lighter sentence by defecting to the police. However, the outcome in which both suspects choose to defect is the worst outcome for the two suspects, compared with the other three outcomes. The two suspects are the players of this prisoner's dilemma. To illustrate this situation as a strategic form game, we use the payoff bimatrix in Table 18.1 to summarize players' payoffs in all the four possible outcomes.

In this payoff bimatrix, the first number in each outcome indicates Player 1's payoff should the corresponding outcome occur, and the second number indicates Player 2's payoff in the corresponding outcome. For example, if both players collude, then every player will receive a payoff of 3. If both players defect, then each will receive a payoff of 1. The goal of every player in the game is to achieve the highest payoff by choosing one of two actions, given the action played by the other player. The payoffs can be ranked in order of desirability from 0 (least desirable) to 4 (most desirable). Desirability reflects the severity of the sentence. A payoff of 0 corresponds to a more severe sentence than a payoff of 4. We can illustrate that in this prisoners' dilemma, only the outcome in which both players choose to defect is a Nash equilibrium. Suppose Player 2 chooses to defect; if Player 1 chooses to defect too, Player 1's payoff will be 1, which is higher than Player 1's payoff of 0 if Player 1 chooses to collude. Applying the same argument to Player 2's strategic

**Table 18.1** The Prisoner's Dilemma as a Strategic Form Game

| | *Player 2 colludes and does not provide information to police* | *Player 2 defects by providing information to police* |
|---|---|---|
| Player 1 colludes and does not provide information to police | 3, 3 | 0, 4 |
| Player 1 defects by providing information to police | 4, 0 | 1, 1 |

NOTE: First number refers to Player 1's payoff; second number refers to Player 2's payoff.

thinking implies that Player 2 should choose to defect when Player 1 defects. In the prisoners' dilemma, regardless of which strategy the other player chooses, it is always best to defect. More specifically, when Player 2 chooses to collude, Player 1 will receive a payoff of 4 by defecting, which is higher than his or her payoff of 3 by colluding. When Player 2 chooses to defect, Player 1 will receive a payoff of 1 by defecting as well, which is also higher than his or her payoff of 0 by colluding. In this case, defection is called a dominant strategy, and collusion is a dominated strategy. The best outcome for the two players is when they collude. However, collusion requires that both players choose dominated strategies.

In the Cournot model, between collusion and Cournot equilibrium outcomes, firms face a situation similar to the two suspects in a prisoners' dilemma game. Although firms have higher profits together, it is not an equilibrium outcome, because every firm has an incentive to produce more than the amount agreed to. We will discuss how this collusive outcome may arise as an equilibrium outcome when firms interact repeatedly over time.

### Bertrand Competition

The term *Bertrand competition* refers to the strategic situation when all the firms choose what prices they would like to charge for their products. When firms' products are homogenous, consumers will purchase from the firms that charge the lowest price. Bertrand competition is also a standard strategic form game. However, the strategic situation faced by each firm is different from that in the Cournot competition. Because only the firms that set the lowest price are able to make a sale, a firm's incentive in Bertrand competition is to cut the price of its product as long as it is still able to make profit. Applying this logic, every firm in the market will continue to slash its prices until there is no profit to make. When all the firms have a common constant marginal cost and no fixed cost, they will set the price at the constant marginal cost, which is also the constant average cost in this simple setting. If all firms set their prices at the average cost and one firm sets its price lower, this firm will make a loss. On the other hand, if a firm sets its price higher than the average cost, this firm will not make any positive sale, and hence, it has zero profit. Therefore, a firm does not have a profitable deviation when the other firms set their prices at the average cost of production. This equilibrium outcome in which all the firms choose the price at the average cost and every firm has zero profit is referred to as the *Bertrand equilibrium.*

## Dynamic Oligopoly Models and Backward Induction

In a dynamic environment, the strategic interactions among firms are more complicated than in a static environment. A dynamic strategic interaction generally concerns an environment that involves sequential decision making by the agents. In a dynamic model, it is necessary to be more specific about when an agent chooses an action, what actions this agent can choose, and more importantly, what this agent knows at the time a choice of action is made. Because of this new element in dynamic strategy interaction, we must make the distinctions between actions and strategies more explicit. We first review how to model a dynamic game, as well as some widely used equilibrium concepts. Then we discuss a number of different dynamic oligopoly models to illustrate how economic agents interact in those dynamic settings.

### Extensive Form Games

Researchers model dynamic strategic interactions using extensive form games. Extensive form games provide all the necessary elements, including the agents or the players of the game, the timing of each player's choice of action, which actions each player may choose, what each player knows when choosing an action, and what each player cares about. Unlike in a strategic form game, the information available to a player when choosing an action is new in a dynamic game. As in static strategic problems, a player will choose an action based on the information available. One common way to summarize information available to a player is to introduce the history of past events up to the point at which the player chooses an action. In a dynamic game, every player needs to make a plan about which action to take after each possible event. A complete plan of action for a player, including one available action after every relevant event, is the player's strategy. In other words, a strategy of a player in a dynamic game is a contingent plan of actions, a plan contingent on the histories available to the player. In a static game situation, there are no different stages of information, so the concept of strategies is the same as that of actions.

As in a static game, any list of strategies (or a strategy profile of one strategy by each player) will induce a complete path of plays, which can be considered as an outcome of the game. When choosing which strategy to play, a player evaluates strategies based on the outcomes that the strategies induce, given the strategies adopted by the other players.

Depending on how well a player knows what has happened, a dynamic game can have either perfect information or imperfect information. Perfect information refers to dynamic strategic situations in which every player knows exactly what has happened, including the actions taken by the players in the past. Many dynamic games in real life have perfect information, such as the game of chess, in which every player technically observes all the past steps. The formulation of extensive form games, however, does not exclude the possibility of some players choosing actions simultaneously, in which case the information structure is imperfect. Technically, we can collect the set of players, players' strategies, and their payoffs to form a

strategic form representation of a dynamic game. However, a strategic form representation of a dynamic game loses its dynamic information structure.

## Nash Equilibrium and Subgame Perfect Nash Equilibrium

We can represent any dynamic game with a strategic form game. Any Nash equilibrium in the strategic form representation of a dynamic game is also a Nash equilibrium in the dynamic game. Because strategic representation of a dynamic game loses its information on dynamic structure of the model, applying the concept of Nash equilibrium in a dynamic game has been criticized because Nash equilibrium may involve some unreasonable behaviors, referred to as noncredible threats. Noncredible threats are the actions that are not one of the best responses given the history and what players will play in the future. Noncredible threats can be part of a Nash equilibrium strategy profile when they occur after histories that are unreachable for the given strategy profile, and hence, they are irrelevant to players' payoffs. However, if those histories were reached, noncredible threats would not be carried out because they are not the best response given how the rest of the game will be played.

Applying the concept of Nash equilibrium in dynamic games oversimplifies the strategic reasoning. In a dynamic game, when any player chooses an action, this player should choose the best action available given the history and how the rest of the game ought to be played when the same reasoning is applied. In dynamic games with a finite horizon, this line of reasoning leads to the method known as backward induction. Backward induction analyzes the last stage of the game first. In games with perfect information, the problem in the last stage of the game is a simple decision problem; the active player simply chooses one of best available actions during the last stage of the game, given how the game has been played. Next, we analyze players' strategic rationale in the second-to-last stage of the game, given how the game has been played and how the game will be played in the last stage of the game.

The problem faced by the active player in the second-to-last stage of the game is a well-defined decision problem induced by what will happen in the last stage of the game after every action of the players in the second-to-last stage. Accordingly, the active player will choose the best action induced by what will happen during the last stage. After identifying what will happen in the last two stages of the game, we then analyze the third-to-last stage and so on, and eventually we will solve a subgame perfect Nash equilibrium of the model. Backward induction eliminates noncredible threats so that any strategy profile resulting from backward induction does not suffer the criticism we discussed for applying the concept of Nash equilibrium in a dynamic game. In a dynamic game of perfect information, strategy profiles resulting from backward induction form a special class of Nash equilibria known as subgame perfect Nash equilibria.

There may not be well-defined stages in a dynamic game with imperfect information, so we cannot directly apply the backward induction technique used for games with perfect information. Instead, researchers first identify all situations in which the rest of the game can be viewed as a dynamic game itself. These situations are called the subgames of the underlying dynamic game. In fact, each stage of a dynamic game with perfect information is a subgame. After identifying all the subgames, we first derive Nash equilibrium outcomes in every subgame that does not include any other subgame—the class of smallest subgames. We then analyze subgames that contain only these smallest subgames, and the Nash equilibria that induce a Nash equilibrium in each of the smallest subgames it contains. Repeating this process leads to a Nash equilibrium of the entire game that induces a Nash equilibrium in any subgame. Accordingly, such a Nash equilibrium is subgame perfect Nash equilibrium. The concept of the subgame perfect Nash equilibrium is a widely adopted refinement of Nash equilibrium in economic research because of its simplicity. Many other refinements of the Nash equilibrium concept have been introduced and studied. The References and Further Readings section provides some useful sources on the details of equilibrium refinements in extensive form games.

## Stackelberg Model

The Stackelberg model refers to situations in which firms in oligopoly markets compete in their output, but firms do not all make their decisions at the same time. Naturally, the underlying strategic situation is a dynamic game. For simplicity, we discuss the Stackelberg model with two firms. One firm, called the quantity leader, chooses and commits to how much to produce. The other firm, called the quantity follower, first observes the amount of output by the quantity leader and then decides how much it will produce. In this subsection, we focus on this simple Stackelberg model and illustrate how various concepts and ideas for dynamic games are applied.

The Stackelberg model is a two-stage dynamic game with perfect information. To the quantity leader, there is a unique history when it makes a decision on how much to produce. Because the quantity follower observes how much the quantity leader has chosen to produce, each output level chosen by the leader is a history to the follower when it makes its decision about how much to produce. This difference sets the Stackelberg model apart from the Cournot model.

To apply the backward induction technique to find the subgame perfect Nash equilibrium in this Stackelberg model, first consider the second stage of the model in which the quantity leader has already committed to its output level. The strategic problem faced by the quantity

follower simplifies to an ordinary monopoly problem with the residual demand, which is the difference between market demand and the amount of output committed to by the quantity leader. Applying exactly the same argument as in the Cournot oligopoly model, we can find the quantity follower's best response to the amount of output by the quantity leader. The follower's best response naturally varies with respect to the quantity leader's output level and can be considered as the Nash equilibrium outcome in the subgame induced by the output quantity committed by the quantity leader.

When the quantity leader decides how much output to produce during the first stage of this Stackelberg model, it should able to anticipate how the quantity follower will respond. Taking the best response of the quantity follower into account, the problem faced by the quantity leader reduces to a decision problem because the amount of output by the quantity follower is uniquely determined by how much the quantity leader produces, and hence, the profit of the quantity leader is also uniquely determined by its output level. Any solution to this induced profit optimization problem for the quantity leader's strategy is a subgame perfect Nash equilibrium of the Stackelberg model. The best response correspondence of the quantity follower is its equilibrium strategy because its best response specifies what the quantity follower should do in all possible subgames in the second stage of the model. The response of the quantity follower to the equilibrium strategy (quantity) of the quantity leader is what would happen in this subgame perfect Nash equilibrium, which is the equilibrium outcome. All other subgames in the second stage will be not reached, according to the equilibrium strategy of the quantity leader.

## Repeated Strategic Interaction

In many economic and other strategic problems, players often interact repeatedly over many time periods. Because the strategic situation during each time period is the same, a repeated game model provides a relatively simple framework to study players' long-term strategic behavior. People change their behavior when they consider the future as well as their immediate well-being. A well-known phenomenon in repeated interaction is the folk theorem; repeated interaction introduces many new equilibrium outcomes.

To illustrate how repeated interaction differs from its static counterpart, we revisit the Cournot model with two firms. Recall that the collusive outcome is not a Nash equilibrium in the static environment, even though both firms have higher profits than in the static Cournot-Nash equilibrium. Collusion cannot be sustained in equilibrium in the static Cournot model because every firm has an incentive to produce more. Nevertheless, when both firms value their future highly enough, collusion is possible when it is highly likely that the two firms will be involved in the same Cournot competition in the future. Collusion can be sustained in a repeated Cournot model by a class of simple strategies known as *grim trigger strategies*. A grim trigger strategy profile calls for each firm to collude at the beginning of the repeated game and to continue to collude as long as both firms have colluded in the past, and as soon as either or both firms deviate, both firms will choose their Cournot-Nash equilibrium strategies forever. Any deviation from collusion triggers the Cournot-Nash equilibrium in the future, which serves as a punishment to the deviating firm. First observe that if a cartel breaks down, it is in each firm's best interest to follow its Cournot-Nash equilibrium strategy. Now the question is what incentive each firm has to continue to collude if they have colluded in the past. Consider the following two possibilities: If a firm deviates away from the collusion outcome, its profit in the current period will be higher. However, such a deviation will trigger the static Cournot-Nash equilibrium forever. On the other hand, if it does not deviate, collusion will continue. Applying the same reasoning in every period, each firm is facing two sequences of profits when deciding whether to continue to collude: collusion forever or a higher profit in the current period followed by a lower profit from Cournot-Nash equilibrium forever. When it is highly likely that the two firms will be involved in the same situation again and again, and when every firm values the future highly enough, a firm will find it beneficial to forgo short-term gain in exchange for collusion in the future. Here we discuss only one particular subgame: perfect Nash equilibrium outcome in the infinitely repeated Cournot model. As matter of fact, all outcomes in which every firm has a profit may arise as an equilibrium outcome in the infinitely repeated Cournot model. This result is known as the folk theorem.

## Conclusion

In this chapter, we took competition among oligopoly firms as a background to discuss a number of key concepts and model ideas in analyzing strategic interactions. We first considered how a perfectly competitive firm, and then a monopolistic firm, behaves, to pave the road for oligopoly problems. The key message is that firms make decisions in response to any change in the economic environment. In the rest of this chapter, we focused on oligopoly markets and analyzed how firms compete in various markets. We started with the Cournot model, in which firms compete in their outputs and make decisions simultaneously. This problem was formalized as a strategic game, and we applied the concept of Nash equilibrium to this model. Borrowing the idea of individual incentives, we have argued the collusive outcome in the Cournot model is not an equilibrium. We also discussed the Bertrand model, in which firms choose their prices simultaneously. We then moved to dynamic strategic interactions. We reviewed the Stackelberg model as

an extensive form game and the concept of a subgame perfect Nash equilibrium. Economics of strategy applies game theory in economics and in business. Here we provided only a brief overview of firms' strategic behaviors. Because of its versatility, game theory has been applied in virtually all areas of economics research and studies.

## References and Further Readings


Aliprantis, C. D., & Charkrabarti, S. K. (2000). *Games and decision making.* New York: Oxford University Press.

Besanko, D., Dranove, D., Shanley, M., & Schaefer, S. (2009). *Economics of strategy.* Hoboken, NJ: Wiley.

Cabral, L. M. B. (2000). *Introduction to industrial organization.* Cambridge: MIT Press.

Dixit, A. K., & Nalebuff, B. J. (1991). *Think strategically: The competitive edge in business, politics, and everyday life.* New York: W. W. Norton.

Dixit, A. K., & Skeath, S. (1999). *Games of strategy.* New York: W. W. Norton.

Friedman, J. (1989). *Oligopoly theory.* New York: Cambridge University Press.

Mendelson, E. (2004). *Introducing game theory and its applications.* New York: Chapman & Hall/CRC.

Osborne, M. J. (2004). *An introduction to game theory.* New York: Oxford University Press.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory.* Cambridge: MIT Press.

Rasmusen, E. (2007). *Games and information: An introduction to game theory* (4th ed.). Malden, MA: Blackwell.

Shy, O. (1996). *Industrial organization, theory and applications.* Cambridge: MIT Press.

# 19

# Transaction Cost Economics

Christopher Ross Bell
*University of North Carolina at Asheville*

Economics is the study of human interactions involving trade. In addition to the costs of production and distribution, every trade has associated with it the costs of the trade itself: of locating trading partners, of negotiating the terms of the trade and adapting those terms as economic conditions change, of monitoring and enforcing the terms of the trade, and so forth. These costs are called *transaction costs,* and traders' attempts to limit them explain why many economic, political, and social institutions have evolved the way they have.

Transaction cost economists seek to understand how transaction costs influence the design and evolution of the institutions used to coordinate economic relationships. Friedrich Hayek (1945) has written of the "marvel" of the spontaneous coordination of the market in a world of flux and change: It aggregates and makes efficient use of information no one person could ever know, it is flexible and capable of responding quickly to changing economic conditions, and it avoids undue concentrations of economic and political power (p. 527). But there is an alternative to market coordination: *hierarchical coordination,* defined as coordination through deliberate choices made by managers or central planners. All modern economic institutions embody varying degrees of these alternative coordination mechanisms.

Though most modern economies rely heavily on market coordination, it is easy to find hierarchical coordination in even the most market-oriented economies: Significant portions of our private sectors are composed of vertically integrated firms, firms in which internal flows of physical materials are coordinated by orders from superiors rather than by independent responses to price signals. Further, families, schools, religious groups, social organizations, nonprofit organizations, and governments all engage in economic activities, and all are coordinated through nonmarket means.

Examples of questions undergraduates can (and have) address using the tools of transaction cost economics include the following: Have the rules guiding trades on eBay changed over time, and if so, have they changed in a manner consistent with the principles of transaction cost economics? Are the rules guiding trades on massively multiplayer online role-playing games consistent with the principles of transaction cost economics? Are there patterns to which services particular universities own and operate themselves and which they contract out? (Who operates your bookstore? Your dining hall? Your library?) Are there patterns to which services hotels provide themselves and which they contract out? Restaurants? Minor-league baseball teams? When services are contracted out, how are the contracts structured? Why do tattoo artists frequently rent space in shops owned by others but always own their own tattoo machines? Are the relationships and patterns of asset ownership found between tattoo artists and the shops from which they rent space similar to those between hairdressers and beauty shops? Mechanics and garages? Doctors and hospitals? Are the similarities and differences consistent with the predictions made by transaction cost theory?

Transaction cost economics is particularly well suited for study by undergraduates. First, it is one of the few economic subdisciplines whose key concepts are better expressed in words than equations. This means that students can read almost its entire corpus—from its classic works to its cutting edge—without the need for mathematics any more advanced than that required for their introductory economics course. If they read the key works in

chronological order, they can watch the economic frontier expand before their eyes, enabling them to understand the often halting process of economic inquiry in a manner unattainable through reading textbooks and later works written with the benefit of hindsight and reflection. Second, many students find transaction cost economics interesting because its study crosses disciplinary boundaries, relaxes assumptions, and explores subjects frequently left unaddressed in other economics courses. Thus, it challenges their views of what economics is, what it is not, and how it relates to other disciplines. Third, the tools and techniques of transaction cost economics are easily applied to real-world situations about which students have firsthand knowledge and interest. This makes it relatively straightforward for them to conduct research projects testing its fundamental propositions.

This chapter provides an introduction to transaction cost economics, with a focus on the classic works in the field. The first section lays the theoretical foundation, describing its origins, defining its key concepts, and deriving its central testable hypothesis. After discussing and illustrating its archetypical application, the make-or-buy decision, this section concludes with a discussion of the broad array of institutional arrangements that lie between spot markets and vertical integration. The second section applies the theory laid out in the first section and examines the empirical evidence supporting it. The focus of the second section is on early applications that highlight the key concepts and principles laid out in the first section and on early tests of the theory that exemplify approaches from which students writing term papers may draw inspiration. Much of the interest in transaction cost economics is policy driven. The third section examines these policy implications. The fourth section summarizes the key points made in the rest of the chapter and suggests directions for future research. This chapter closes with suggestions for further reading.

## Transaction Cost Theory

### Origins

Ronald Coase's (1937) "The Nature of the Firm" is the key work in transaction cost economics. It is a remarkable treatise for many reasons, not the least of which is that it was conceived when its author was just 21 years of age (Coase, 1988a). Coase became interested in the questions of vertical and lateral integration while studying for his bachelor of commerce degree at the London School of Economics. Why, he wondered, if the invisible hand of the price system was as wonderful as his professors made it out to be, did firms intentionally replace the spontaneous coordination of the market with the deliberate coordination of the manager? Coase spent the academic year 1931–1932 exploring this intriguing question in a series of visits to U.S. businesses and industrial plants—the happy coincidence of the necessity for an additional year of study after completing the courses required for his degree with his receipt of a scholarship that allowed for travel abroad (Coase, 1988a).

Coase (1988a) first laid out his answer to his question in a letter written in October 1932, describing a lecture he had just given at the Dundee School of Economics and Commerce (p. 4). Following reflection on what he had observed and discussed with plant managers, Coase noted that there were costs to using the price system, costs we now know as transaction costs. Coase reasoned that firms emerge when the transaction costs associated with coordinating particular exchanges using markets exceeded those of coordinating the same exchanges using managers. In the absence of transaction costs, Coase (1988c) wrote, "The firm has no purpose" (p. 34).

Coase published his explanation in 1937. As he was to write in 1972, for close to 35 years, "The Nature of the Firm" was "much cited and little used" (Coase, 1988b, p. 23). One explanation may be that Coase's observation that vertically integrated firms exist because the advantages of replacing the costs of market exchange with those of command and control is fundamentally tautological: Any and every substitution of one means for organizing economic activity for another can be explained as the outcome of a desire to reduce transaction costs. Because Coase's theory was not fleshed out enough to predict in advance which exchanges would be the most likely candidates for vertical integration, it explained both everything and nothing (Fischer, 1977, p. 322; Williamson, 1979, p. 233).

Oliver Williamson was one of the first to recognize that Coase's fundamental insight represented the beginning of a research agenda, not the end. Beginning in the early 1970s, Williamson, his students, and others who have been inspired by his and Coase's work have produced a flourishing body of research that has grown into a rich and powerful set of intellectual tools, tools amenable to those hallmarks of science, prediction, and test. Especially important have been the works that formed the basis of Williamson's *Markets and Hierarchies: Analysis and Antitrust Implications* (1975), *The Economic Institutions of Capitalism* (1985), and *The Mechanisms of Governance* (1996).

### Key Concepts

One of the more frustrating aspects of taking up a new subject is learning its jargon. Six concepts central to transaction cost economics are *transaction, transaction cost, governance structure, bounded rationality, opportunism,* and *asset specificity.*

*Transaction.* Trade requires transactions, that is, transfers of goods or services across "technologically separable interface[s]" (Williamson, 1981, p. 552). Note that while trade between businesses, individuals and businesses, governments and businesses, and so forth necessarily results in transactions, transfers of goods or services across

technologically separable interfaces are almost certainly at least as common within businesses, families, social organizations, and governments. Thus, just as the transfer of a six-pack of beer from your local convenience store's shelves to your refrigerator represents a transaction, so does the transfer of aluminum cans from a brewer's aluminum-can-manufacturing operations to its brewing operations.

*Transaction cost.* Transactions have costs associated with them: locating trading partners, negotiating the terms of trade, writing explicit contracts prior to the commencement of trade, monitoring compliance with the contracts and enforcing them as trade progresses, adapting the terms of trade to address changing circumstances, and observing, quantifying, and processing the information required to do all of the above. It is noteworthy that these transaction costs differ from those economists ordinarily emphasize—production costs—in that it is assumed "the former can be varied by a change in the mode of resource allocation, while the latter depend only on technology and tastes" (Arrow, 1969, p. 60).

*Governance structure.* What Arrow referred to as "mode[s] of resource allocation" (1969, p. 60) are now called governance structures. The term was coined by Williamson in his 1979 paper, "Transaction-Cost Economics: The Governance of Contractual Relations." Governance structures are examples of *institutions,* defined as sets of laws, rules, customs, and norms that guide human behavior (North, 1994). Thomas Palay (1984) provided an unusually clear definition of the term, writing that "governance structure" is "a shorthand expression for the institutional framework in which contracts are initiated, negotiated, monitored, adapted, enforced, and terminated" (p. 265). Examples of governance structures include spot markets (your purchase of gasoline at a convenience store while driving to spring break) and vertical integration (an oil company's production of the crude oil it will later refine into gasoline). Although the reasons different governance structures generate different transaction costs will be dealt with in greater detail later, in defining the rules of the game (customers are required to pay in advance for the gas they pump, and convenience stores are required to maintain accurate meters on their gas pumps), well-designed governance structures create incentives for transactors to cooperate with each other.

*Bounded rationality.* As a step toward understanding why the choice of governance structure matters, one must abandon the fiction of the all-knowing economic man. Transaction cost economists assume that economic decision makers are subject to bounded rationality; that is, while their actions are "intendedly rational," cognitive limitations render them "only limitedly so" (Simon, 1957, p. xxiv).

The cognitive limitations that give rise to bounded rationality come in many forms. First and foremost, we live in an incredibly complex and ever-changing world. Even if we could fully comprehend all those parts of the world critical to the economic decisions we face, what we know today may no longer be relevant when we wake up in a new world tomorrow. Worse, what we will need to know tomorrow may well be unknowable today. Governance structures differ both in regard to the amount and depth of the information individuals need and in their ability to adapt to changing information. Remember what Hayek (1945) found marvelous about the governance structure we call the market: its ability to aggregate and make efficient use of information no one person could ever know, including its ability to aggregate and make efficient use of new information.

Second, even if we could know everything there is to know about those parts of the world most important to us—past, present, and future—we would still be hampered by our ability to make sense of what we know. Herbert Simon (1972) emphasized the difficulties faced by people who must process large amounts of information. Consider, Simon (1972) suggested, writing out a decision tree for a chess game and then using this tree to decide which moves to take. Although doing so would reveal the optimal move in any given situation, no one, not even a grand master, could construct and make efficacious use of such a tree. Instead, chess players husband their limited cognitive resources by relying on relatively simple strategies that they know from experience will guide them to favorable—though not necessarily optimal—outcomes (pp. 165–171). Similarly, governance structures differ with regard to their demands on economic actors' ability to process information; assigning activities to alternative governance structures in a transaction-cost-economizing way requires strict attention to the differences in cognitive demands alternative governance structures make.

John McManus (1975), Douglass North (1981), and Yoram Barzel (1982) emphasize yet another cognitive limitation with implications for institutional design: the difficulty, and at times impossibility, of quantifying the information needed to make, monitor, and enforce agreements. Barzel and Roy Kenney and Benjamin Klein (1983) offer the diamond sights arranged by the DeBeers group, in which dealers are offered mixed lots of presorted diamonds on a take it or leave it basis, as an example of an economic institution that has evolved to reduce the costs incurred when cognitively limited human beings must quantify information.

*Opportunism.* Opportunism is self-interest taken to its limit. To be opportunistic is to act in ways that combine "self-interest seeking with guile" (Williamson, 1975, p. 26). As Williamson (1985) wrote, "This includes but is scarcely limited to more blatant forms, such as lying, stealing, and cheating. Opportunism more often involves subtle forms of deceit." These subtle forms include, "the incomplete or distorted disclosure of information, especially calculated efforts to mislead, distort, disguise, obfuscate, or otherwise confuse" (p. 47). Recognition of our proclivity for opportunism deepens our understanding as to why the choice

of governance structure matters. Just as the assignment of activities to alternative governance structures requires strict attention to the differences in cognitive demands alternative structures make, so too does the assignment require strict attention to the abilities of alternative governance structures to discourage opportunism.

*Asset specificity.* Some transactions require investments in durable (i.e., long-lived) assets that have little or no value except in support of the transaction for which the investment has been made. For example, auto racks—the railcars railroads use to transport finished automobiles—are useful only for transporting automobiles and are built in custom configurations to carry specific models produced by specific manufacturers (Palay, 1984, p. 269). Williamson (1979) used the term *idiosyncratic transaction* to refer to transactions supported by transaction-specific assets; a number of synonyms are used to describe the investments themselves, including *idiosyncratic investments* (and assets) and *transaction-specific investments* (and assets).

There are many sources of asset specificity. Investments in auto racks exemplify the term *physical asset specificity:* They represent an investment in physical capital in a custom configuration with little value in any but the transaction for which they are designed. Investments in transaction-specific human capital (e.g., the value of the time a mechanical engineer spends mastering the technical details of products unique to his or her employer and of value to no one else) represent the term *human asset specificity.* Other sources include *site specificity* (e.g., an investment in physical capital located at a specific site that is costly to relocate and has little value except when located in close proximity to the transaction it supports), *dedicated asset specificity* (e.g., an investment in physical capital that increases capacity that is not needed except in support of the transaction it was made to support), and *brand-name capital specificity* (e.g., an investment developing, promoting, and maintaining the goodwill associated with a specific branded good or service).

Two important points about transaction-specific investments cannot be overemphasized. First, the investments of interest are indeed transaction specific. Consider the investment in time and money a student makes learning to be a mechanical engineer. This investment is specific in the sense that it trains the student to be a mechanical engineer, but it is not transaction specific because there are many alternative transactions requiring general mechanical engineering skills. Second, transaction-specific investments are not made for the sake of making transaction-specific investments but because they stimulate final sales or reduce production costs (Williamson, 1981, p. 558). Thus, it may make sense to make transaction-specific investments in dies designed to stamp uniquely appealing auto bodies because doing so stimulates sales or to make transaction-specific investments in custom auto racks because doing so reduces the cost of transporting automobiles from the manufacturer to the dealer, but it would not make sense to do so in the absence of benefits like these.

## A Testable Hypothesis

The central testable hypothesis of transaction cost economics has come to be known as the *discriminating alignment hypothesis.* It rests on three propositions:

*Proposition 1: The transaction costs associated with particular transactions depend on the governance structures within which the transactions take place.* This is a restatement of Arrow's (1969) observation that transaction costs "can be varied by a change in the mode of resource allocation" (p. 60). The reasons for believing this proposition will be discussed in greater detail later, but remember that governance structures can be viewed as defining the rules of the game, and as James Buchanan (1975) had so often argued, changing the rules of the game can change outcomes—including the transaction costs associated with the play of the game.

*Proposition 2: Transactions tend to be matched to alternative governance structures in ways that reduce the sum of production and transaction costs.* The matching may be due to conscious choice or it may be the outcome of evolutionary processes (e.g., survival in a competitive world). This proposition is a restatement of Coase's (1937) seminal insight that vertically integrated firms exist because of the advantages of replacing the costs of market exchange with those of managerial coordination.

As noted earlier, however, Proposition 2 is, by itself, tautological: Any and every substitution of one governance structure for another can be explained as the outcome of a desire to reduce transaction costs. Williamson in "The Vertical Integration of Production: Market Failure Considerations" (1971) and "Transaction-Cost Economics: The Governance of Contractual Relations" (1979), and Benjamin Klein, Robert Crawford, and Armen Alchian in "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process" (1978) broke the tautological nature of Coase's argument with proposition three:

*Proposition 3: Alternative governance structures differ in their abilities to economize on particular types of transaction costs; likewise, exchanges differ in their proclivity to generate particular types of transaction costs.* This proposition implies that researchers should, after identifying the appropriate characteristics of exchanges and of governance structures, be able to predict ex ante which types of exchanges will be matched to which types of governance structures. The ability to do so allows Coase's proposition to be tested.

## The Archetypical Application: The Make-or-Buy Decision

The classic application of transaction cost economics is to the make-or-buy decision: Should a firm make an input itself, or should it buy it from an outside vendor? The discriminating alignment hypothesis suggests a three-step

approach. First, assess the production and transaction-cost-reducing advantages and disadvantages of the competing governance structures. Second, identify the critical dimensions with respect to which transactions differ from each other in their proclivity to generate particular types of costs. Third, choose the governance structure whose cost-reducing advantages and disadvantages are the best match to the types of transaction costs the transaction in question is likely to generate.

Consider as Williamson did (1979, p. 245) the potential advantages to a firm of buying the input from an outside vendor. Looking first at the transaction costs, if the market for the input is competitive, the interaction of demand and supply will determine the terms of trade, present and future. This reduces the transaction costs associated with negotiating the initial terms and adapting them to changing economic conditions. The existence of alternative buyers and sellers, should either party try to take advantage of the other, encourages both to do what they have agreed to do. This reduces the transaction costs associated with enforcing the terms of trade. Finally, larger firms are more difficult to manage than smaller firms because of the difficulties bureaucracies have preserving the high-powered incentives provided by markets (Williamson, 1985, chap. 6). All else equal, by choosing to buy rather than to make, the firm will be a little smaller and easier to manage.

There are two reasons buying the input from an outside vendor may lead to lower production costs. First, if by meeting the needs of multiple customers, an outside vendor can achieve greater economies of scale, the vendor's production costs will fall, a cost savings that in a competitive market will be passed on to the vendor's customers. Second, if demand fluctuates over time, by combining production for customers whose demand for the input is less than perfectly correlated, the vendor can reduce the variability in his or her production runs in the same way an investor holding a diverse portfolio can reduce the variability in his or her returns. This too lowers production costs.

Given the many advantages of buying from outside vendors, why would a firm ever want to make the input itself? Like markets, hierarchies have their own distinctive strengths and weaknesses. With vertical integration, both sides to the transaction become part of a common team. This encourages the pursuit of a single goal—the success of the team—rather than the opportunistic pursuit of the separate goals of the parties on either side of the transaction. Further, when conflict arises, higher-level managers have access to financial, market, and technical information that can help settle disputes in ways they do not when negotiating with independent suppliers or customers. Finally, if a higher-level manager's access to lower-level information is not by itself sufficient to settle a dispute, the higher-level manager can simply declare the terms on which the dispute will be resolved. Such management by fiat is in general a more effective and less costly means of conflict resolution than third-party arbitration or litigation.

By reducing the incentives for opportunistic behavior and by increasing the options for dispute resolution, vertical integration can be especially attractive for transactions that require frequent adaptations to changing economic conditions and in which competition from rival traders cannot be relied on to discourage opportunistic behavior. Rather than write an initial contract spelling out the terms of trade for every future contingency (the better to avoid opportunistic renegotiation of the terms of trade as the future becomes the present), the vertically integrated firm can replace extensive contingent contracting and its demands on the transactors' limited cognitive powers with adaptive, sequential decision making.

Against these strengths, the weaknesses of vertical integration must be weighed. First, by vertically integrating, the firm gives up the potential production cost savings available to an input supplier meeting the needs of multiple customers. Second, as already discussed, vertically integrated firms are larger firms, and all else equal, larger firms are more difficult (read more expensive) to manage than smaller firms.

Thus, in answering the question, should we make or should we buy? the firm must weigh the distinctive strengths and weakness of vertical integration against the distinctive strengths and weakness of market procurement. Ordinarily, we would expect the weight of the comparisons to favor market procurement (Williamson, 1979). Note, however, that the advantages of market procurement hinge on the related assumptions that the market for the input is competitive and that the outside vendor can realize cost savings by meeting the combined needs of multiple customers. To the extent that these conditions are not met, the advantages of buying over making decrease, the disadvantages increase, and alternatives to market contracting begin to look more attractive. An important reason these conditions may not be met is that the transaction in question is supported by transaction-specific investments.

It is time to turn to the second step of the discriminating alignment hypothesis: What are the critical dimensions with respect to which transactions differ from each other? Williamson (1979) identified three transaction characteristics that in combination make vertical integration more likely: the frequency with which a transaction recurs, the degree of uncertainty associated with the transaction, and the specificity of the assets supporting the transaction.

Frequency matters for several reasons, one of which is that greater frequency implies that the extra costs associated with creating more complex governance structures can be spread over a greater number of transactions. This reduces the marginal cost of bringing the transaction under the direct control of the firm's managers, which, when combined with uncertainty and asset specificity, makes vertical integration more likely.

Uncertainty is important because it makes contracting over markets more complex and increases the likelihood that both sides of a transaction will find it in their interest to renegotiate the transaction's terms as economic conditions change. As an example of both uncertainty and the incentives firms have to adapt their contracts as conditions change, consider the relationship between a commercial

aircraft manufacturer designing a new plane and the vendor from which it buys engines. An airliner's engines are among its most important components. Many aspects of the design and performance of the new design will depend on the specific characteristics of the engines used, engines that are designed in tandem with the new plane so that both can take advantage of the most recent technological advances. Because the technology used is so new, however, the actual characteristics of the engines—including the cost to produce them—may not be known with certainty until the engine is in production. The fate of both the aircraft manufacturer and its engine vendor can be tied to their willingness to renegotiate the terms of their relationship as technological and market conditions unfold. This example is based on the experiences of the Boeing Company and the Lockheed Aircraft Manufacturing Company and the difficulties they experienced bringing the Boeing 747 and the Lockheed L-1011 to market; Lockheed's struggles brought it to the brink of bankruptcy. Boeing is today experiencing similar troubles bringing their 787 Dreamliner to market, due in part to the uncertainties associated with building the first large commercial airliner whose wings and fuselage are constructed out of fibers held together with epoxies rather than aluminum held together with fasteners.

It is worth noting that the importance of uncertainty in the make-or-buy decision is directly related to bounded rationality: If not for their cognitive limitations, traders should, at least in theory, be able to write contracts specifying the terms of trade under every possible future state of the world. Given their cognitive limits, however, both the expense and the ultimate impossibility of conceiving and enumerating all possible future states of the world prevent them from doing this. One alternative is to write contracts that are intentionally incomplete, then adapt them to the conditions that emerge as the future becomes the present. However, in the presence of the third transaction characteristic that makes vertical integration more likely—asset specificity—incomplete contracting opens the door to opportunistic behavior.

Asset specificity can lead to contracting problems because investments in transaction-specific assets lock the parties to transactions into bilateral (i.e., one-to-one) relationships in which each party is dependent on the other and competitive forces cannot be depended on to quickly, easily, and impartially determine the terms of trade. Because of this dependency and lack of competition, an environment ripe for opportunistic gamesmanship is created each time the terms of trade are renegotiated: If the first party owns the transaction-specific asset, the second party can threaten to withdraw from the relationship if the first refuses to renegotiate the terms in the second's favor. The gamesmanship and the transacting problems it can lead to have become known as *the hold-up problem.*

Klein et al. (1978) were the first to describe the hold-up problem in detail, introducing the concept of the *appropriable quasi rent* to do so. An appropriable quasi rent is the difference between the net revenue ownership of an asset generates for its owner when deployed in support of its current transaction at the current terms of trade and the net revenue it generates when deployed in support of its next best alternative transaction. In the absence of asset specificity, the appropriable quasi rent is zero: If the first party refuses to renegotiate and the second party withdraws from the relationship, the first can redeploy his or her assets with a third party at the original terms of trade and no loss in net revenue. At the other extreme, if the asset owned by the first party is uniquely suited to the transaction with the second party, there will be no third party to which the first can turn, and the entire difference between the net revenue at the current terms of trade and the salvage value of the asset is potentially in play.

Note that even in cases in which markets are competitive before investments in transaction-specific assets are made, after the investment, the parties are locked into the noncompetitive bilateral relationship that can lead to the hold-up problem. This shift from a competitive to a noncompetitive situation after the specific investments are made is called *the fundamental transformation.* When answering the make-or-buy question, firms must consider the possibility that the fundamental transformation will occur, and if so, whether they can devise a nonmarket governance structure that will encourage cooperative behavior without placing unrealistic demands on their limited cognitive capacities and forcing them to give up the production cost savings associated with market provision.

Two final points before moving on to an illustration: First, transaction cost economics is an explicitly comparative institutional endeavor. Williamson (1975, p. 130) reminded us that all institutions have associated with them characteristic strengths and weaknesses. Finding that market procurement in a particular situation has hazards associated with it is not the same as finding that vertical integration is necessarily the superior alternative. The challenge for businesses (and the economists studying their decisions) is to match troublesome transactions to imperfect institutions in a discriminating way.

Second, the matching of transactions characterized by frequency, uncertainty, and asset specificity to more complex governance structures—up to and including vertical integration—rests on a crucial assumption. This assumption was first clearly and unambiguously stated by Klein et al. (1978), who wrote,

> The crucial assumption underlying the analysis of this paper is that, as assets become more specific and more appropriable quasi-rents are created (and therefore the possible gains from opportunistic behavior increases), the costs of contracting will generally increase more than the costs of vertical integration. (p. 298)

Combined with Williamson's (1979) insight that asset specificity alone is not sufficient to tilt the scales away from market contracting and that frequency and uncertainty are

required to create the conditions in which asset specificity matters, it follows that the greater the degree to which frequency, uncertainty, and asset specificity are combined, the more likely it is that the answer to the question, should we make or should we buy? will be to make.

## An Illustration: Adam Smith's Pin Factory

The prevalence of opportunism combined with bounded rationality creates interesting problems for our political, economic, and social institutions to overcome. Consider, for example, the trading hazards faced by people working together to produce a simple product. In his *Wealth of Nations,* Adam Smith (1776/1974) described the manufacture of pins circa 1776:

> One man draws out the wire, another straights it, a third cuts it, a fourth points it, a fifth grinds it at the top for receiving the head; to make the head requires three distinct operations; to put it on is a peculiar business, to whiten the pins is another; it is even a trade by itself to put them into the paper; and the important business of making a pin is, in this manner, divided into about eighteen distinct operations. (p. 110)

Imagine the problems faced by any two of Smith's pin makers if the manufacture of pins is organized along entrepreneurial lines, that is, if the governance structure used to coordinate flows of semifinished pins rests on repeated market exchanges between independent entrepreneurs at each of the 18 separable stages in a pin's manufacture. Because the pin makers are only boundedly rational, they will have difficulty foreseeing and then writing a contract covering all the possible contingencies over which disagreements could occur in the future. Because each has reason to suspect the other may on occasion be opportunistic, each will be concerned that his or her trading partner may on occasion try to take advantage of him or her. This means that neither can place complete faith in a contract that says, "I will always negotiate in good faith when unforeseen contingencies arise and always provide accurate and timely information." Further, their physical proximity locks them into a bilateral trading situation; this reduces the role competitive forces can play in guaranteeing harmonious exchange. So they may look for alternatives to market organization, alternatives that realign the incentives they face to act at cross-purposes. In particular, they may choose to form a vertically integrated firm.

Note something important: By choosing a different governance structure, the formerly independent entrepreneurs have altered their incentives to cooperate with each other. That is, through careful manipulation of their trading environment, they have sought to make mutually desired outcomes (harmonious adaptation to changing circumstances) more likely. Doing so may well entail some costs: bias in information flows, additional information-processing costs, bureaucratic inertia, and so forth, but in an imperfect world, some transaction costs are unavoidable; the trick is to find that governance structure that economizes the transaction costs associated with a particular exchange. In Smith's pin factory, this has been accomplished by (in Arrow's, 1969, words), replacing "the costs of buying and selling on the market by the costs of intrafirm transfers" (p. 48).

## Beyond the Make-or-Buy Decision: Governance as a Continuum

Although the make-or-buy decision may be both the original and the archetypical application of transaction cost economics, a moment's reflection on the incredible diversity of trading arrangements in our economy leads to the conclusion that most of the action is neither in the sorts of spot markets that dominate principles of economics texts nor in vertical integration. From the beginning of the revival of interest in Coase's work, transaction cost economists have devoted much of their attention to governance structures that lie between spot markets and vertical integration. Williamson (1979) identified four general types of governance structures, ranging from market governance (or classical contracting) at one end of a rough continuum to unified governance (or vertical integration) at the other. In between are trilateral governance (or neoclassical contracting) and bilateral governance (or relational contracting).

In *trilateral governance,* moderate to highly specific investments may be made—creating an incentive for the parties to work together because punishing opportunism by leaving a relationship is costly—but the transactions are so infrequent that the costs of creating a specialized governance structure exceed the benefits. In such cases, the parties to a transaction may rely on a third party to serve as an arbitrator to encourage cooperative behavior and settle disputes. Most construction projects, for example, are managed by a general contractor who hires, coordinates the actions, and settles disputes between the various subcontractors and between the client and the subcontractors.

*Bilateral governance* is arguably the most interesting case. It occurs in situations that combine frequency with moderate amounts of uncertainty and asset specificity. It also occurs in cases that combine significant amounts of frequency, uncertainty, and asset specificity with significant production cost savings realized when the demands of multiple buyers are aggregated. In such situations, highly idiosyncratic long-term contractual, social, and cultural relationships may be crafted to ensure cooperative behavior. For example, bilateral governance structures often include contractual provisions—like the requirements that buyers pay for a minimum quantity of a good, service, or input even if they choose not to accept delivery of the entire minimum order (take-or-pay requirements)—that are hard to square with traditional economic models but make sense when viewed as means to align incentives and

promote cooperative behavior in long-term relationships that might otherwise be problematic.

## Applications and Evidence

A considerable body of evidence exists in support of transaction cost economics' central testable hypothesis: As transaction frequency, uncertainty, and, especially, asset specificity increase (the independent variables), the governance structures used to support transactions become increasingly complex, moving out the governance continuum toward relational contracting and vertical integration (the dependent variable). This body of evidence, which includes case studies, econometric studies, and combinations of the two, overwhelmingly supports Williamson's (1996) assertion that "transaction cost economics is an empirical success story" (p.55). "Indeed," wrote Paul Joskow (2008), "it is hard to find many other areas in industrial organization where there is such an abundance of empirical work supporting a theory of the firm or market structure" (p. 16).

The empirical research on transaction cost economics has been summarized numerous times. Two works in particular stand out: "Empirical Research in Transaction Cost Economics: A Review and Assessment" (Shelanski & Klein, 1995) and "The Make-Or-Buy Decision: Lessons From Empirical Studies" (Klein, 2008). Rather than summarize these summaries, the focus of this section will be on a handful of case studies and empirical tests that highlight key concepts and exemplify approaches from which students writing term papers may draw inspiration.

Palay's (1984) "Comparative Institutional Economics: The Governance of Rail Freight Contracting" has inspired many student papers. There are four reasons for this. First, Palay did an excellent job translating the often idiosyncratic language in which transaction cost theory was developed into language students can understand. Note Palay's definition of governance structure and illustration of physical asset specificity from the theory section of this chapter. Second, he finds ways to break down the theory's key concepts—many of which are both abstract and complex—into smaller, more understandable, measurable pieces. An example of this is his identification of five characteristics of governance structures (how agreements are enforced, how they are adjusted, the types of adjustments made, whether information necessary for long-term planning is exchanged, and whether information necessary for structural planning is exchanged), each of which can be evaluated on scales running from least to most complex (i.e., from what one would expect to see in a spot market to what one would expect to see in relational contracting). Third, the strategy Palay employed to test the relationship between asset specificity and governance—creating a series of tables, one for each of his five governance-structure characteristics, in which the columns represent increasing degrees of governance complexity, the rows represent increasing degrees to which transaction-specific investments have been made, and the numbers in the individual cells represent the number of transactions characterized by particular combinations of governance complexity and asset—is simple, intuitive, and does not require a course in economic statistics to apply: Look to see if the expected pattern between asset specificity and governance emerges (the greatest numbers of observations should form diagonals down the tables, with cells in the upper left, center, and lower right of each table most full). And fourth, collecting the raw data can be fun—it is done by conducting a series of interviews about specific transactions, asking questions that allow the interviewer to characterize each transaction in terms of governance-structure complexity and investments in transaction-specific investments, uncertainty, and frequency. Parents, employers, college administrators, and local business owners can all be interviewed, depending on the student's research project.

One criticism some students have of the classic works in transaction cost economics is their focus on heavy industry: commercial aircraft producers and their subcontractors (Bell, 1982), rail freight contracting (Palay, 1984), natural gas producers and pipelines (Masten & Crocker, 1985), coal and electric utilities (Joskow, 1987), steel producers and source of iron ore (Mullin & Mullin, 1997), and so forth. There are good reasons for reading these works. The significance of Palay's work was discussed above, and that of Joseph Mullin and Wallace Mullin's will be discussed in the section of this chapter on policy implications. The extensive reliance on subcontracting in the production of commercial aircraft described by Bell reminds us that high frequency, uncertainty, and transaction-specific investment need not lead to vertical integration if the benefits of achieving economies of scale and smoothing production by aggregating less than perfectly correlated demand in different end markets are high. Scott Masten and Keith Crocker (1985) did a great job describing the sort of odd contractual provisions (take-or-pay requirements, in their case) that are difficult to explain when one views the world as a series of spot markets but are the hallmarks of relational contracting. Of the many virtues of Joskow's (1987) article, one is that his discussions of site and physical asset specificity (mine mouth power plants built next to coal mines and power plants specially built to efficiently burn coal with the specific metallurgical properties of the coal unique to a colocated mine) help remind us that asset specificity is not chosen for its own sake but rather for its demand-enhancing or production-cost-reducing benefits. But are the only applications of transaction cost economics to heavy industry? Thankfully, the answer is no.

Examples of applications and empirical tests of transaction cost economic outside heavy industry include academic tenure (McPherson & Winston, 1983), fiscal federalism (Bell, 1988, 1989), marketing and distribution in the carbonated beverage industry (Muris, Scheffman, & Spiller, 1992), franchising in the fast food industry

(Kaufman & Lafontaine, 1994), and even marriage (Hamilton, 1999). Applications by students have added to the variety, as demonstrated by the list of term paper topics in the introductory section of this chapter.

## Policy Implications

The roots of the revival of interest in what we now call transaction cost economics were policy driven. Both Arrow (1969) and Williamson (1967) conducted research on government contracting in general and defense contracting in particular just as the revival was getting under way, exploring many of the same issues Williamson in particular would soon explore in greater depth in the context of commercial contracting. Just as an awareness of the issues raised by bounded rationality, opportunism, the fundamental transformation, and the potential for contractual provisions in long-term contracts that realign incentives in a way that promotes cooperative adaptations to changing economic conditions are important to businesses transacting with other businesses, they are important to governments transacting with businesses. The issues involved should play an important part in the ongoing and increasingly important debate over the future of the U.S. health care system. But historically, interest in the policy implications of transaction cost economics has been strongest in the area of antitrust.

Significant portions of two of the key early works on transaction cost economics, Williamson's *Markets and Hierarchies* (1975) and his *The Economic Institutions of Capitalism* (1985), are devoted to the antitrust implications of transaction cost economics, and many of the key early papers were published in law journals. This is no coincidence. Before the advent of transaction cost analysis, the default assumption in the antitrust community was that odd contractual provisions, that is, provisions that differed much from what one might expect to see in a spot market, were suspicious. Transaction cost economics provided an alternative explanation: Rather than mechanisms for the enhancement and exploitation of market power, these odd contractual provisions promoted economic efficiency. Examples cited by Williamson (1985) in the first chapter of *The Economic Institutions of Capitalism* included "customer and territorial restrictions, tie-ins, block booking, franchising, vertical integration, and the like" (p. 19).

The motivation for much of the interest in transaction cost economics has thus been to distinguish between contractual provisions most likely used to enhance market power (which is illegal) and those most likely used to promote economic efficiency (which is not illegal, can be used as a defense in an antitrust case, and in general should be encouraged). Mullin and Mullin's (1997) "United States Steel's Acquisition of the Great Northern Ore Properties: Vertical Foreclosure of Efficient Contractual Governance?" provided an interesting illustration. Mullin and Mullin revisited a historic antitrust case, one in which U.S. Steel stood accused of vertical foreclosure (creating an illegal barrier to entry into a market by eliminating a potential rival's access to a critical input) through their negotiation in 1906 of a long-term lease of ore properties owned by the Great Northern Railway. This relationship, called the Hill Ore Lease, is particularly interesting because its structural characteristics are consistent with both the market power and transaction-cost-reducing explanations for its existence. The two interpretations have different implications, however, for the postcontractual profits and stock prices of U.S. Steel's largest customers at the turn of the last century—the railroads. Mullin and Mullin used ordinary least squares, the capital asset pricing model, and railroad stock prices to see with which interpretation returns on railroad stocks over the period the lease was in effect were consistent. Because these returns were abnormally high, Mullin and Mullin concluded that the primary effect of the lease was the promotion of economic efficiency, not market power.

## Conclusion

Transaction cost economists believe that either through decision makers' conscious choices or by doing what is necessary to survive in a competitive world, economic institutions evolve in ways that reduce the sum of production-plus-transaction costs. When the exchanges between the various technologically separable steps required to bring a good or service to market are simple, spot markets tend to do this best. However, as the exchanges become more complex, the transaction costs associated with the use of spot markets begin to rise, and alternatives to spot-market contracting become more appealing. One alternative is vertical integration. Just as spot markets have transaction costs associated with their use, however, so does vertical integration. For this reason, comparisons of the costs and benefits of the various institutional alternatives generally result in pairings of particular transactions to governance structures that lie between the extremes of spot-market contracting and vertical integration.

To test transaction cost theory, one must identify those factors that lead to greater transaction complexity. Theory predicts that complexity increases as the combination of transaction frequency, uncertainty, and the presence of transaction-specific investments increases, all else equal, including the production-cost benefits associated with capturing economies of scale and smoothing production by aggregating demand in less than perfectly correlated end markets. This theory has been tested using case studies, econometric studies, and combinations of the two. The overwhelming weight of the evidence is in support of transaction cost theory.

A number of important reminders emerge from these tests and applications, including the frequency with which real-world governance stops short of vertical integration and the consequent prevalence of relational contracting

and its odd contractual provisions. These provisions have historically been viewed with suspicion by antitrust authorities. With the advent of transaction cost reasoning, they are seen in a new light—as means to reduce transaction costs and promote efficient adaptations to changing economic conditions rather than as mechanisms by which to exploit and enhance market power.

## References and Further Readings


Arrow, K. J. (1969). The organization of economic activity: Issues pertinent to the choice of market versus nonmarket allocation. In *The analysis and evaluation of public expenditures: The PPB system* (Vol. 1, pp. 47–64). Washington, DC: 91st Congress, Joint Economic Committee.

Barzel, Y. (1982). Measurement costs and the organization of markets. *Journal of Law and Economics, 25,* 27–48.

Bell, C. R. (1982). The theory of the firm: Transaction, production, and system cost considerations. In *Transaction Cost Economics Workshop Working Paper.* Philadelphia: University of Pennsylvania, Economics Department.

Bell, C. R. (1988). The assignment of fiscal responsibility in a federal state: An empirical assessment. *National Tax Journal, 41,* 191–207.

Bell, C. R. (1989). Between anarchy and Leviathan: A note on the design of federal states. *Journal of Public Economics, 39,* 207–221.

Buchanan, J. (1975). A contractarian paradigm for applying economic theory. *American Economic Review, 65,* 225–230.

Coase, R. H. (1937). The nature of the firm. *Economica, New Series, 4,* 386–405.

Coase, R. H. (1988a). The nature of the firm: Origin. *Journal of Law, Economics, & Organization, 4,* 3–17.

Coase, R. H. (1988b). The nature of the firm: Meaning. *Journal of Law, Economics, & Organization, 4,* 19–32.

Coase, R. H. (1988c). The nature of the firm: Influence. *Journal of Law, Economics, & Organization, 4,* 33–47.

Fischer, S. (1977). Long-term contracting, sticky prices, and monetary policy: Comment. *Journal of Monetary Economics, 3,* 317–323.

Hamilton, G. (1999). Property rights and transaction costs in marriage: Evidence from prenuptial contracts. *Journal of Economic History, 59,* 68–103.

Hayek, F. A. (1945). The use of knowledge in society. *American Economic Review, 35,* 519–530.

Joskow, P. L. (1987). Contract duration and relationship-specific investments: Empirical evidence from coal markets. *American Economic Review, 77,* 168–185.

Joskow, P. L. (2008). Introduction to the new institutional economics: A report card. In E. Brousseau & J. M. Glachant (Eds.), *New institutional economics: A guidebook* (pp. 1–19). New York: Oxford University Press.

Kaufman, P. J., & Lafontaine, F. (1994). Costs of control: The source of economic rents for McDonald's franchisees. *Journal of Law and Economics, 37,* 417–453.

Kenney, R. W., & Klein, B. (1983). The economics of block booking. *Journal of Law and Economics, 26,* 497–540.

Klein, B., Crawford, R. G., & Alchian, A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics, 21,* 297–326.

Klein, P. G. (2008). The make-or-buy decision: Lessons from empirical studies. In C. Menard & M. M. Shirley (Eds.), *Handbook of new institutional economics* (pp. 435–464). Berlin, Germany: Springer-Verlag.

Masten, S. E., & Crocker, K. J. (1985). Efficient adaptation in long-term contracts: Take-or-pay provisions for natural gas. *American Economic Review, 75,* 1083–1093.

McManus, J. (1975). The costs of alternative economic organizations. *Canadian Journal of Economics, 8,* 334–350.

McPherson, M. S., & Winston, G. C. (1983). The economics of academic tenure: A relational perspective. *Journal of Economic Behavior & Organization, 4,* 163–184.

Mullin, J. C., & Mullin, W. P. (1997). United States Steel's acquisition of the Great Northern Ore properties: Vertical foreclosure or efficient contractual governance? *Journal of Law, Economics, and Organization, 13,* 74–100.

Muris, T. J., Scheffman, D. T., & Spiller, P. T. (1992). Strategy and transaction costs: The organization of distribution in the carbonated soft drink industry. *Journal of Economics and Management Strategy, 1,* 83–128.

North, D. C. (1981). *Structure and change in economic history.* New York: W. W. Norton.

North, D. C. (1994). Economic performance through time. *American Economic Review, 84,* 359–368.

Palay, T. M. (1984). Comparative institutional economics: The governance of rail freight contracting. *Journal of Legal Studies, 13,* 265–287.

Shelanski, H. A., & Klein, P. G. (1995). Empirical research in transaction cost economics: A review and assessment. *Journal of Law, Economics and Organization, 11,* 335–361.

Simon, H. A. (1957). *Administrative behavior* (2nd ed.). New York: Macmillan.

Simon, H. A. (1972). Theories of bounded rationality. In C. B. McGuire & R. Radner (Eds.), *Decision and organization* (pp. 161–176). New York: Elsevier.

Smith, A. (1974). *The wealth of nations.* New York: Penguin Books. (Original work published 1776)

Williamson, O. E. (1967). The economics of defense contracting: Incentives and performance. In R. N. McKean (Ed.), *Issues in defense economics* (pp. 218–256). New York: National Bureau of Economic Research.

Williamson, O. E. (1971). The vertical integration of production: Market failure considerations. *American Economic Review, 61,* 112–123.

Williamson, O. E. (1975). *Markets and hierarchies: Analysis and antitrust implications.* New York: Free Press.

Williamson, O. E. (1979). Transaction cost economics: The governance of contractual relations. *Journal of Law and Economics, 22,* 233–261.

Williamson, O. E. (1981). The economics of organization: The transaction cost approach. *American Journal of Sociology, 87*(3), 548–577.

Williamson, O. E. (1985). *The economic institutions of capitalism.* New York: Free Press.

Williamson, O. E. (1996). *The mechanisms of governance.* New York: Oxford University Press.

# 20

# Asset Pricing Models

Peter Nyberg
*Helsinki School of Economics*

One of the central questions in financial economics is how to assign a correct value to an asset that provides a stream of uncertain future cash flows. At the most general level, the solution is simple: The price of the asset today equals the expected discounted value of its future payoffs. This discounting—the assignment of a (usually) lower present value to future cash flows—reflects the time and risk dimensions of the payoffs offered by the asset. First, most investors value a dollar received today more than they value a dollar received tomorrow. Consequently, they are willing to pay less than one dollar today for a dollar that they will receive in the future. This is the time dimension of asset pricing. Second, because investors dislike uncertainty and demand a compensation for bearing extra risk, the rate at which the payoffs are discounted should reflect how risky these payoffs are. Riskier payoffs are discounted more heavily than payoffs that are less risky. This is the risk dimension of asset pricing. The variation in riskiness across assets gives rise to variation in their expected returns: Because of higher discounting, assets with a riskier payoff pattern should have a lower price and hence, provide higher expected returns than assets that are otherwise similar but have less risky payoffs. This higher expected return is the reward that investors demand so that they are willing to include the risky asset in their portfolios. In equilibrium, then, the rate of return on an asset should be aligned with its riskiness in order for demand by risk-averse investors to meet supply. Hence, the interplay between risk and return—the search for the fundamental sources of risk and the quantification of this risk—lies at the heart of asset pricing research.

Asset pricing theory provides guidance on how assets optimally should be priced and how investors optimally should invest under some assumptions specified in the theoretical models. On the other side of the coin, asset pricing research also aims to understand how assets actually are priced on real financial markets. Thus, there is a close interplay between theory and practice. For investors operating in the financial markets, it is of fundamental interest to know the fair price of any given financial asset and the return they can expect to get from their risky investments. Furthermore, using the tools of modern portfolio theory, investors can quantify the trade-off between risk and return to construct optimal portfolios that suit their own risk preferences. Hence, asset pricing theory has vast practical implications and applications. Considering this, it is not surprising that the advent of modern portfolio theory and asset pricing models revolutionized Wall Street and the field of investment management. See, for example, Peter Bernstein (2005), who provides a historical account of the interplay between theory and practice of financial economics.

This chapter proceeds as follows: First, the general framework for understanding risk and return is reviewed. Then, some specific asset pricing models, the capital asset pricing model, the intertemporal capital asset pricing model, and the consumption-based asset pricing model, are described. This is followed by a brief review of the empirical performance of the asset pricing models. This chapter concludes with some practical implications, some guidelines for future research, and a summary section. A list of further readings is also provided.

## Understanding Risk and Return

We start the analysis by assuming a single-period economy, starting at Date 0 and ending at Date 1. For simplicity, we refer to Date 0 as today and Date 1 as tomorrow. The

analysis can easily be extended to multiple periods, but the simplified valuation framework considered here is sufficient for deriving the general asset pricing results.

A representative[1] investor living in our simplified economy is faced with the problem of choosing how much of his or her wealth he or she will allocate for consumption today and how much he or she will save and allocate for consumption tomorrow. The allocation of consumption between today and tomorrow is facilitated by the existence of an asset market. An asset is a financial contract that requires a cash outlay today from the buyer and that provides a payoff to the buyer in the future. The payoff provided by the asset can then be used for consumption—that is, for buying some units of a consumption good in the future.

The investor makes his or her investment decisions today, at Date 0, and receives a payoff from his or her investments tomorrow, at Date 1. The fact that the investment is made today and the payoff is obtained in the future reflects the time dimension of asset pricing. However, to generate some relevant predictions for asset prices, we must also incorporate uncertainty about the future into the analysis. This can easily be done by assuming that there are different states of nature that can occur tomorrow. It is also assumed that the payoffs from the assets are state contingent in the sense that the payoff for a given asset might vary over the possible states of nature, but once we know what state will occur, we also know with certainty the payoff that the asset will provide. We also assume that we can assign probabilities to these mutually exclusive states of nature so that these probabilities for the different states sum to 1. For example, with two possible future states of nature, the economy tomorrow might be in a recession with a probability of $\pi$ or in a boom with a probability of $(1 - \pi)$. Risk in this framework is then captured by the fact that the investor, making his or her investment decisions today, will not know with certainty the state of nature and, consequently, the associated payoffs from his or her assets that will be realized tomorrow.

Having described the basic framework, the investor's consumption-portfolio problem can now be stated.

Start by assuming that there are $S$ different states of nature that can occur tomorrow, indexed by $s = 1, 2 \ldots, S$. Associated with each of these states is the probability that the state will occur, $\pi_1, \pi_2, \ldots, \pi_S$, with these probabilities summing to 1. The investor now has to choose how much he or she wants to consume today, $c_0$, and how much he or she would want to consume in each of the mutually exclusive states of nature tomorrow, $c_1, c_2, \ldots, c_S$.

Now, the question becomes how the investor can decide on an optimal allocation of consumption. Clearly, we need a decision rule with respect to which optimality can be defined. One such decision rule is given by the expected utility theory, developed by John von Neumann and Oskar Morgenstern (1947). Thus, we assume that the investor is choosing his or her consumption levels today and in the alternative states of nature tomorrow so that his or her expected utility from these consumption levels is maximized. Furthermore, we also assume that the investor always prefers more consumption to less consumption, and that he or she is risk averse. Technically, these two conditions imply that the utility function $u(c)$ that describes the preferences of the investor always increases when the consumption level grows larger (the function has a positive first derivative) but that the rate of increase in utility for a fixed increase in consumption gets smaller as the level of consumption grows larger (negative second derivative).

The investor's choice problem can now be stated as follows. He or she should choose the consumption levels $c_0$, $c_1, c_2, \ldots, c_S$ so that his or her expected utility

$$U(C) = u(c_0) + \theta \sum_{s\ 1}^{s} \pi_s u(C_s), \tag{1}$$

subject to the budget constraint

$$c_0 + \sum_{s\ 1}^{s} p_s c_s = W, \tag{2}$$

is maximized. In Equation 1, $\theta$ is a time-discount factor, usually assumed to be less than one ($\theta < 1$). This parameter captures the time preferences of the investor: The utility that the investor gains from a given amount of consumption tomorrow is less than the utility that he or she would gain from the same level of consumption today. In Equation 2, $W$ denotes the investor's current wealth. The variable $p_s$ is the current price (measured today) of one unit of consumption in the possible state of nature $s$ that can occur tomorrow. Alternatively, $p_s$ can be viewed as the current price of a security that offers as a payoff a claim to one unit of consumption if state $s$ is realized and nothing if any other state is realized. In the literature, such securities are often referred to as *primitive securities*,[2] and $p_s$ is the associated state price for the given state $s$. Furthermore, the price of one unit of consumption today is normalized to be 1. Consequently, the budget restriction states that the current wealth of the investor is used for consumption today and for buying contingent claims to consumption in the future.

The optimization problem presented in Equations 1 and 2 can be solved by using the method of Lagrange multipliers. For the purposes of this chapter, it suffices to show the first-order conditions of the optimization—that is, the conditions that must hold when consumption is allocated optimally so that the expected utility is maximized. These are

$$u'(c_0) = \lambda, \tag{3}$$

and

$$\theta \pi_s u'(c_s) = \lambda p_s, \tag{4}$$

for all $s = 1, 2 \ldots, S$, where $\lambda$ is the Lagrange multiplier and $u'(\bullet)$ denotes the first derivative of the utility function.

Equations 3 and 4 offer important intuition about the interpretation of the state prices $p_s$. We shall later see that these state prices are important determinants of the prices and expected returns on all financial assets. Note that by combining the two equations, the state prices can be written as

$$p_s = \theta \pi_s \frac{u'(c_s)}{u'(c_0)}. \tag{5}$$

The relationship presented above shows that there are three major factors that determine the state prices. First, we have the time preference of the investor captured by the effect of the time-discount factor $\theta$. Second, the probability of a state occurring, $\pi_s$, also affects the state prices. The smaller the probability that the claim to a unit of consumption in a future state can be used, the less valuable the claim will be. Finally, the third determinant—the ratio of marginal utilities between state-contingent consumption tomorrow and consumption today—turns out to be the most important one for asset pricing purposes. Recall that we assumed that the investor was risk averse. This is captured by the fact that the second derivative of his or her utility function is negative, implying that the first derivative, the marginal utility, is a decreasing function of consumption. Consequently, states with high aggregate consumption are characterized by low marginal utilities, whereas states with low aggregate consumption are characterized by high marginal utilities. This implies that state prices, everything else equal, are low for states where aggregate consumption is high and that they are high for states where aggregate consumption is low. A claim to one unit of consumption is more valuable and has a higher price when this claim can be used in a state of economy where resources are scarce and aggregate consumption is low.

## The Pricing of Individual Assets and the Risk-Free Rate

We now turn the discussion to the valuation of individual assets. As mentioned earlier, in this framework, we view a financial asset as an instrument that offers payoffs that are contingent on the state of nature that will occur. Consider an asset that provides state contingent payoffs $D_{i,s}$, where $i$ is an asset-specific subscript and $s$ denotes the state. For example, if there are two possible states of nature, recession (State 1) and boom (State 2), then the payoffs offered by some asset indexed by 1 could be $D_{1,1} = 3$ and $D_{1,2} = 6$. This implies that the asset offers a payoff of three claims to consumption in a recession and six claims in a boom. The asset can now be valued using the state prices. Note that the state price $p_s$ is the price today for a payoff of one unit received in state $s$ tomorrow. Thus, the current value of the payoff offered by the asset in state $s$ must be equal to the state price (value of a unit payoff) multiplied by the units of payoff that the asset offers in the given state. Consequently, any financial asset can be assigned a price today simply by taking a sum of the current values of its payoffs across all states, or

$$P_i = \sum_{s=1}^{s} p_s D_{i,s}, \tag{6}$$

where the capital letter $P_i$ denotes the price of the asset today, $p_s$ is the state price for state $s$, and $D_{i,s}$ is the state-contingent payoff provided by the asset. Using Equation 6, it is also straightforward to calculate the expected return on the asset. The state-contingent rate of return on any given asset is defined as the payoff of the asset in a given state tomorrow divided by its current price. Denote this state-contingent payoff return by $r_{i,s}$, that is,

$$r_{i,s} = \left( \frac{D_{i,s}}{P_i} - 1 \right). \tag{7}$$

The expected return on the asset can then be calculated as a probability-weighted outcome of the state-contingent rate of returns. In symbols,

$$E(r_i) = \frac{\sum_{s=1}^{s} \pi_s D_{i,s}}{P_i} - 1 = \sum_{s=1}^{s} \pi_s r_{i,s}. \tag{8}$$

The asset pricing framework described here makes it also easy to characterize a risk-free asset. A risk-free asset is an asset that gives the same state-contingent payoff in all possible states of the economy. Such an asset is risk free because it promises a payoff that does not vary over the states. If we normalize the payoff from such an asset to be one unit in all possible states tomorrow, the current price of the risk free asset is given by

$$P_{rf} = \sum_{s=1}^{s} p_s, \tag{9}$$

and the risk-free rate of return can be written as the reciprocal of the sum of the state prices,

$$r_{rf} = \frac{1}{P_{rf}} - 1 = \frac{1}{\sum_{s=1}^{s} p_s} - 1. \tag{10}$$

## Stochastic Discount Factors and Risk Premiums

Having reviewed the general framework for pricing assets, we now turn to the important question of what determines the risk premiums on different assets. One of the important research themes in empirical asset pricing is to understand why some assets provide higher average returns than some other assets. Intuitively, we would expect that the differences in returns between assets should be driven by the riskiness of their payoffs. Because investors are risk averse, they are willing to invest in riskier assets only if the risks of these assets are compensated through a higher expected return. But what do we mean when we say that one asset is riskier than another asset, and how is the expected return on the asset connected to its riskiness? This section uses the general framework

presented in the previous sections to formalize the relationship between risk and return on financial assets.

We start the analysis by rewriting the general pricing Equation 6 slightly.

$$P_i = \sum_{s\_1}^{s} p_s D_{i,.s} = \sum_{s\_1}^{s} \pi_s \frac{p_s}{\pi_s} D_{i,s} = \sum_{s\_1}^{s} \pi_s M_s D_{i,s} = E(MD_i), \quad (11)$$

$$\text{where } M_s \equiv \frac{p_s}{\pi_s} = \theta \frac{u'(c_s)}{u'(c_0)}. \quad (12)$$

In the second equality of Equation 11, we have simply multiplied and divided the pricing expression with $\pi_s$, the probability of a given state occurring tomorrow. This enables us to write the pricing formula in the form of an expectation, illustrated in the last equality of Equation 11. The price of an asset is given by a probability-weighted average of the product across all states between the probability-deflated state price ($M_s$) and the state-contingent payoff of the asset.

The variable $M$ plays a central role in modern asset pricing. Depending on the application in which it is used, it is often called a stochastic discount factor, state-price deflator, state-price density, or pricing kernel in the literature. In the remainder of this text, we refer to $M$ as a stochastic discount factor because it appears to be the most common name used. The definition presented in Equation 12 shows that the stochastic discount factor carries the interpretation of an intertemporal marginal rate of substitution. It obtains high values in states of the economy where consumption is low and marginal utilities are high.

Note that Equation 11 can be expressed in terms of returns instead of prices. To see this, divide both sides of the equation with the current price of the asset.

$$\frac{P_i}{P_i} = E\left(M \frac{D_i}{P_i}\right) \quad (13)$$
$$1 = E(MR_i),$$

where $R_i = (1 + r_i)$ is the gross return on the asset.

Equation 13 lends itself to further manipulation. In what follows, we are going to use a well-known result from basic statistical theory. Recall that the covariance, a measure of the tendency of two random variables $X$ and $Y$ to move together, can be written as $\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$. Rewriting this expression slightly, we obtain $E(XY) = E(X)E(Y) + \text{Cov}(X,Y)$. Using this relationship on the variables involved in Equation 13 yields the result presented below.

$$1 = E(MR_i) = E(M)E(R_i) + \text{Cov}(M, R_i) \quad (14)$$

Using the expression for the risk-free rate presented in Equation 10 together with the definition of the stochastic discount factor implies that the gross risk-free rate can be written as $R_f = 1/E(M)$. Equation 14 then implies that

$$E(R_i) - R_f = -R_f \text{Cov}(M, R_i) \quad (15)$$

Alternatively, starting from $P_i = E(MD_i)$ and using the same covariance decomposition as above leads to an equivalent expression for prices.

$$P_i = \frac{E(D_i)}{R_f} + \text{Cov}(M, D_i). \quad (16)$$

It is important to understand the intuition underlying Equations 15 and 16. We see that the risk premium on an asset, defined as the difference between the asset's expected return and the return on a risk-free investment, is given by the negative of the covariance between the asset's return and the stochastic discount factor. Assets that deliver high returns when the stochastic discount factor is low and low returns when the stochastic discount factor is high have a negative covariance with the stochastic discount factor. Such assets demand a positive risk premium, and the stronger the negative association between the returns and the stochastic discount factor, the higher will the required risk premium be.

A higher expected return, everything else equal, implies a lower current price for the asset. Equation 16 demonstrates that the current price of the asset can be decomposed into two parts. The first part is the expected payoff that the asset delivers tomorrow, discounted at the risk-free rate. If investors did not care about risk, or if the payoffs of the asset were risk free, then the current price of the asset would be given by the first term on the right-hand side of Equation 16. The second term, the covariance of the asset's payoff with the stochastic discount factor, is a correction for the riskiness involved in the asset's payoffs. The more negative this covariance is, the lower the current price of the asset and the higher its expected return will be.

Recall from Equation 12 that the stochastic discount factor $M$ has the interpretation as the intertemporal marginal rate of substitution. This allows us to give more structure to the argument above. Because $u(\bullet)$ is assumed to be a concave function, the marginal utility $u'(\bullet)$ is high when the quantity of the variable from which the investor derives utility is low. In bad states of the economy, characterized by low aggregate consumption or low aggregate wealth, marginal utilities tend to be high, leading to high values of the stochastic discount factor. These are exactly the states of nature where an investor values an extra payoff more than he or she would value it in states where marginal utilities are lower. Everything else equal, assets that provide high payoffs in states of high marginal utility—that is, assets that have a positive covariance with the stochastic discount factor, will thus be more attractive to the investor than assets that have a negative covariance with the stochastic discount factor and thus tend to provide these high payoffs in states where marginal utilities are low. For demand to meet supply, these latter assets, due to their unattractive payoff pattern, must provide higher expected returns (Equation 15) and have lower prices (Equation 16) so that investors are willing to include them in their portfolios. In equilibrium, the higher expected return on these assets will offset their unattractive payoff pattern, and

investors will, at the margin, be indifferent in their choice between different assets.

### Systematic and Unsystematic Risk

Maybe somewhat surprisingly, the expression for the risk premium presented in Equation 15 does not involve the variance of the asset's return. Instead, it is the covariance of the asset's return with the stochastic discount factor that determines the risk premium. This covariance term captures the amount of systematic risk involved in the asset's payoffs: Only the part of the asset return that is correlated with the stochastic discount factor matters for the investor. Conversely, the part of an asset's return variance that is unconnected to the stochastic discount factor represents unsystematic risk that does not affect the risk premium on the asset.

We can again understand this prediction by recalling that the stochastic discount factor is a function of the investor's marginal utilities in the different states of nature. An investor cares about how the asset behaves in good and bad states of the economy, where the marginal utility of the investor reflects the state of the economy. An asset with a payoff variation that is completely unconnected to the variation in marginal utilities behaves like a risk-free asset: Its payoff pattern is not directly connected to the well-being of the investor in the different states of nature tomorrow.

The distinction between systematic and unsystematic risk carries over to all asset pricing models. In the general asset pricing framework studied here, where marginal utility is a function of consumption, an asset that is uncorrelated with the stochastic discount factor is also uncorrelated with the investor's consumption stream. That the asset's payoffs are uncorrelated with consumption implies that if an investor adds such an asset to his or her portfolio, the payoffs from the asset will not, at the margin, affect the variance of his or her total consumption stream. In models in which the marginal utilities are a function of the returns on well-diversified portfolios, such as the capital asset pricing model described in detail later on, the same argument carries through. In this case, adding to a portfolio a slight amount of an asset that is uncorrelated with the portfolio returns will not, at the margin, affect the variance of the investor's wealth. The distinction between systematic and unsystematic risk implies that assets cannot be evaluated in isolation, but the evaluation must always be based on how the asset, in combination with other assets, affects the overall well-being of the investor.

## Description of Specific Asset Pricing Models

Having reviewed the general framework for pricing assets, the discussion now turns to some specific asset pricing models. The first model to be reviewed is the capital asset pricing model. Then, the intertemporal capital asset pricing model and the consumption-based asset pricing model will be briefly described.

### The Capital Asset Pricing Model (CAPM)

The CAPM, derived by William Sharpe (1964), among others, can be viewed as the first formal model that quantifies the relationship between expected return and systematic risk for financial assets. The model lays out the equilibrium implications of the portfolio selection models of Harry Markowitz (1952) and James Tobin (1958) that are briefly reviewed below.

Markowitz (1952) was the first to formalize how diversification—the combination of several risky assets in a portfolio—affects the expected return and variance of the portfolio. Because individual asset returns are not perfectly correlated with each other, the standard deviation (the square root of variance) of a portfolio consisting of individual assets is less than the weighted average of the individual assets' standard deviations. The weights here are taken to be the relative proportions of the total value of the portfolio that are invested in each one of the assets. On the other hand, the portfolio's expected return is simply a weighted average of the expected returns on the individual assets. Thus, by diversifying, one can reduce the risk of a portfolio without changing its expected return.

One of the key elements of Markowitz's analysis is to characterize the portfolios that for each given level of expected return have the lowest variance. Figure 20.1 shows a graphical depiction of the investment opportunity set available for an investor in a mean-standard deviation space.

The dots inside the hyperbolic region are individual assets, each of them characterized by its mean and variance. For each level of expected portfolio return, the minimum variance that can be obtained by combining these assets into portfolios is represented by the hyperbola in the figure. This is the minimum-variance frontier. If investors have mean-variance preferences[3]—that is, they want to maximize their portfolio returns but they dislike variance because it constitutes risk—they will choose portfolios on the efficient part of the frontier. The efficient frontier is the part of the frontier that lies above the global minimum variance portfolio that is situated at the left-most part of the frontier. Investors with a high risk aversion will choose mean-variance efficient portfolios that have a relatively low expected return and low standard deviation. Investors with less risk aversion, who are willing to take on more risk to get a higher expected return, will choose optimal portfolios that lie further up on the efficient frontier.

Tobin (1958) extends Markowitz's analysis by adding a risk-free asset to the investment opportunity set. The inclusion of the risk-free asset leads to a concept known as the two fund separation theorem: For an investor with mean-variance preferences, it is optimal to divide the investable wealth between the risk-free asset and the

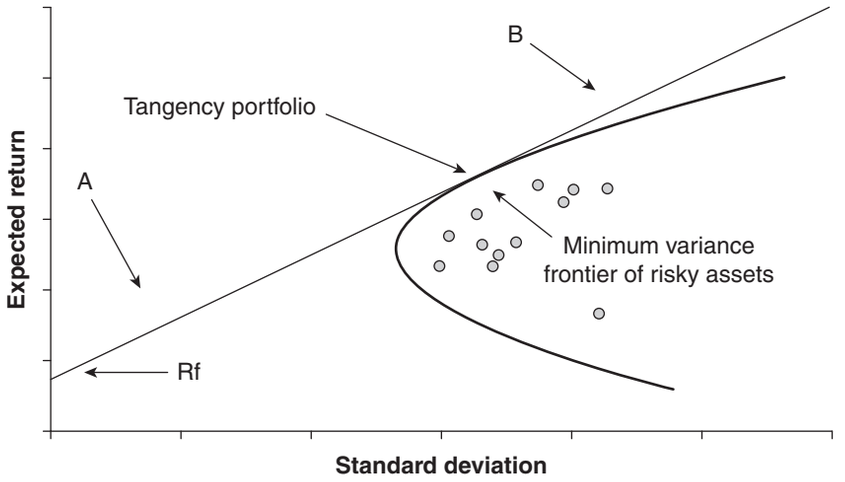


**Figure 20.1** Mean-Variance Opportunity Set

tangency portfolio, located at the point where the line starting from the risk-free asset intersects the efficient frontier. An investor can move up and down on this capital allocation line by varying the weights that he or she invests in the risk-free asset and the tangency portfolio. For example, investor A, who is more risk averse, invests some fraction of his or her wealth in the risk-free asset and the remaining fraction in the tangency portfolio. Investor B, who is willing to take more risk, can borrow money (putting a negative weight on the risk-free asset) and invest more than 100% in the tangency portfolio. Every point on the capital allocation line dominates all other investments in the sense that it is not possible to obtain a lower variance for a given expected return using any other investment strategy.

The CAPM describes an equilibrium in a world where all the investors follow the portfolio advice given above. If there exists a risk-free rate that is the same for all investors, if each investor has an identical holding period, if they all are mean-variance optimizers, and if they all have the same expectations about the probability distribution of asset returns, then they will all identify the same tangency portfolio as their optimal risky portfolio. This tangency portfolio, chosen by all investors, will thus be equal to the market portfolio[4] of risky assets.

While there are many ways to derive the relationship between risk and expected return for individual assets implied by the CAPM, it suffices to state that such a relationship can be obtained by solving a simple portfolio optimization problem. Suppose there are $n$ individual assets and a risk-free asset. The aim is to construct an efficient portfolio by minimizing variance subject to a given target return $E(r_p)$.

$$\begin{aligned} &\min_{w_i} \; Var(r_p) = \sum_{i\ 1}^{n} \sum_{k\ 1}^{n} w_i w_k \mathrm{Cov}(r_i, r_k) \\ &\text{subject to} \sum_{i\ 1}^{n} w_i E(r_i) + (1 - \sum_{i\ 1}^{n} w_i) r_f = E(r_p). \end{aligned} \tag{17}$$

In the equations above, $w_i$ is the weight invested in risky asset $i$, $E(r_i)$ is the asset's expected return, and $\mathrm{Cov}(r_i, r_k)$ is the covariance between assets $i$ and $k$. The first equation is simply the variance of a portfolio of risky assets. This variance is to be minimized by choosing the optimal weights for the individual risky assets. The second equation is a constraint stating that the linear combination of the risky portfolio's expected return and the risk-free rate must equal some target return $r_p$ that lies on the capital allocation line.

By using the Lagrange multiplier method for solving the optimization problem presented in Equation 17 and rearranging terms, one obtains an expression for the expected return on individual assets.

$$E(r_i) = r_f + \beta_{i,\,p}(E\,(r_p) - r_f), \tag{18}$$

where

$$\beta_{i,\,p} = \frac{\mathrm{Cov}(r_i, r_p)}{\mathrm{var}(r_p)}. \tag{19}$$

For each chosen level of expected return, the optimization carried out in Equation 17 gives a portfolio variance that plots on the capital allocation line in Figure 20.1. Thus, the relationship presented in Equation 18 must hold when $E(r_p)$ is the return on any portfolio on the capital allocation line. Setting the weight of the investable wealth invested in the risk-free asset to zero, implying that

$$\sum_{i\ 1}^{n} w_i = 1 \tag{20}$$

gives the expected return and variance on the tangency portfolio. Because of the assumptions of the CAPM, the tangency portfolio is the optimal risky portfolio identified by all investors. Consequently, the tangency portfolio will be equal to the market portfolio. This implies that the relationship in Equation 18 must hold for the market portfolio as well, and we can set $r_p$ to be equal to $r_m$, the return on the market portfolio, in the equations above.

Equation 18, known as the security market line (SML), is the central prediction of the CAPM. It states that the expected return on a security can be divided into two components. The first component is the risk-free rate of return. It corresponds to the rate of return that investors require to switch consumption from this period to the next period. The second component is a correction for risk. The expected risk premium on an asset is given by the product between the asset's beta, $\beta_{i,p}$, and the expected market risk premium $(E(r_p) - r_f)$.

The beta parameter measures the asset's sensitivity to movements in the market portfolio. Assets with a high beta have returns that move in close connection with the returns on the market portfolio. Conversely, assets with a low beta have returns that follow the market portfolio less closely. In line with the general framework studied earlier, we can view the risk premium implied by the CAPM as a compensation for how the asset behaves in the different states of the economy. Assets with a high beta offer high payoffs when the marketwide return is high and low payoffs when the market is doing badly. According to Equation 18, investors require a high-risk premium from these assets. Assets with a low, or even negative, beta enable the investor to diversify away some of the risks connected to market movements. Hence, these assets are more desired, and a smaller risk premium is required from them.

One of the fundamental insights offered by the CAPM is that only a part of the total risk of an individual asset, where total risk is measured by the standard deviation of the asset's returns, should be compensated with a risk premium. This follows because the total risk can be decomposed into a systematic and an unsystematic part. The unsystematic risk of a security, the part of the return variation that is uncorrelated with market movements, can be diversified away in large portfolios. On the other hand, under the assumptions of the CAPM, the systematic risk of an asset is related to its covariance with the market portfolio. This covariance is the part of an asset's total risk that the investor cannot diversify away. Only the systematic risk that investors are forced to bear should be rewarded with higher expected returns in market equilibrium. Hence, in market equilibrium, investors should not be rewarded for holding risks that they can easily diversify away.

### *Roll's Critique*

Note that in the pricing Equation 18, the return on the factor portfolio is denoted by a subscript $p$, instead of using the subscript m that refers to the market portfolio. This choice of notation is intentional. In fact, the pricing Equation 18 holds for any tangency portfolio. To state this in more general terms, a necessary and sufficient condition for finding a perfect relationship between beta and average returns is that the factor portfolio $p$ against which betas are measured is mean-variance efficient—that is, that the factor portfolio's average return and standard deviation plot on the hyperbolic region in Figure 20.1. Such a relationship between betas and expected returns can thus be derived without any need for economic motivations.

It can be argued that the only economic content of the CAPM is that it ex ante identifies that the market portfolio should be mean-variance efficient. The main result of the CAPM, the relationship between beta and expected returns, then follows from the mean-variance efficiency of the market portfolio.

Richard Roll (1977) argues that this has implications for empirical tests of the CAPM. Viewed in this light, a test of the CAPM is a test of the mean-variance efficiency of a market portfolio proxy that we use in our empirical tests. Of course, in every sample there will always exist efficient portfolios. If the proxy happens to be efficient, we obtain a perfect fit between betas and average returns—if it is inefficient, there might be no fit at all. However, even if we should by chance happen to use an efficient stock market proxy, there is no way of knowing whether it corresponds to the true market portfolio.

The only real test of the CAPM would be to test the efficiency of the true market portfolio. However, the true market portfolio is unobservable. It should, in theory, contain all the wealth in the economy and not be limited only to stocks, as is often assumed in the empirical tests. Thus, the CAPM may never be a testable model.

## The Intertemporal CAPM

The original CAPM is a one-period model: It is assumed that investors are trying to maximize the utility of their end-of-period wealth and that they make their portfolio choices without any consideration for what might happen in the next period. This assumption is relaxed in the intertemporal CAPM (ICAPM) derived by Robert Merton (1973). By assuming that investors can trade continuously in time, Merton shows that the expected returns on assets are driven not only by their covariance with the market portfolio but also by the assets' capability to hedge against shifts in future investments opportunities.

A simplified example can illustrate the thrust of Merton's argument. Assume that there is time variation in expected stock market returns and that these returns are predictable to at least some extent. Such time variation could, for example, reflect changes in marketwide risk premiums. If the current prediction were that the stock market returns would be low in the next period, a sufficiently risk-averse investor would like to hedge his or her position against the deterioration in investment opportunities that this implies. One way to hedge the reinvestment risk is to own assets that provide high returns when future stock market returns are predicted to be low. In that case, the investor will have more wealth to allocate to stock market investments at the beginning of the next period, and it is more likely that he or she will achieve some target level of wealth or consumption at the end of the period, in spite of the lower return on the stock market during the period. If the marketwide intertemporal hedging demands are

strong enough, investors will have a high demand for assets that hedge the reinvestment risk—assets that do well when there is an unfavorable shift in future investment opportunities—and this will drive up the price of these assets and decrease their expected returns.

Variables that forecast the conditional distribution of asset returns, or more generally, changes in the investment opportunity set, are called *state variables* in the literature. The ICAPM predicts that the market beta is not enough to describe the expected returns on assets, but we also need to take into account how the returns on the asset covary with the state variables. The betas measured against the state variables indicate the asset's ability to hedge against shifts in investment opportunities. The end result is a multifactor model, where the state variables work as additional factors on top of the market factor.

### Consumption-Based CAPM

Recall the section of this survey that described the general framework for understanding risk and return. The section demonstrated that the consumption–investment problem faced by the representative investor leads to many important asset pricing implications. In fact, as emphasized by John Cochrane (2005), all of the asset pricing models surveyed in this chapter can be seen as specializations of the consumption-based framework. Thus, it is interesting to see how the model can be specified to empirically relevant cases.

The theoretical development of the consumption-based capital asset pricing model (CCAPM) is accredited to Mark Rubinstein (1976), Robert Lucas (1978), and Douglas Breeden (1979). The CCAPM can be derived by considering the intertemporal maximization problem of an investor who can freely, at time $t$, trade asset $i$ that gives a payoff $D_{i,t+1}$ in the next period. The first-order condition for his or her choice is

$$P_{i,t}u'(c_t) = E_t[\theta u'(c_{t+1})D_{i,t+1}]. \tag{21}$$

The left-hand side of Equation 21 is the utility loss faced by the investor if he or she buys the asset and thus forgoes some of his or her consumption today. The right-hand side is the expected utility increase in the next period that he or she gains from his or her increased consumption owing to the extra payoff he or she get will from the asset. At the optimum, the right- and left-hand sides of the equation should be equal. If the investor, at the margin, gained more (discounted) expected utility from his or her increased consumption tomorrow than the utility he or she loses from his or her forgone consumption today, it would be beneficial for the investor to invest more in the asset and move some more of his or her consumption to the next period.

Rearranging the first-order condition slightly, we see that prices are given by

$$P_{i,t} = E_t\left[\theta \frac{u'(c_{t+1})}{u'(c_t)} D_{i,t+1}\right], \tag{22}$$

whereas, in correspondence to the arguments made earlier, returns obey

$$1 = E_t\left[\theta \frac{u'(c_{t+1})}{u'(c_t)} R_{i,t+1}\right]. \tag{23}$$

These results above, and the intuition behind them, are, of course, equivalent to results of Equations 11 and 13, which were derived using the state-preference framework. Furthermore, by specifying a utility function that investors are trying to maximize and using time-series data on consumption and asset returns, Equation 23 is also applicable for empirical testing.

Breeden (1979) derives a version of the CCAPM where the expected return on an asset is linearly related to a single beta. This beta is the sensitivity to the asset's return measured against the growth rate in aggregate real consumption. Thus, in Breeden's CCAPM, consumption growth replaces the return on the market portfolio that was the relevant factor in the CAPM. Furthermore, Breeden also demonstrates that in an intertemporal framework, the single consumption beta captures all the implications of Merton's (1973) multibeta model. This provides a significant simplification of intertemporal asset pricing theory: Whereas an application of the ICAPM requires an identification of all the relevant state variables describing the evolution of the investment opportunity set, the CCAPM needs as an input only the aggregate growth rate of consumption.

## The Empirical Performance of Asset Pricing Models

The previous sections have reviewed the relevant theory and described some asset pricing models. The next step is to assess whether the predictions made by the asset pricing models can match the properties that are observed in real asset return data. The empirical literature on asset pricing models is enormous, and only a brief review of some of the most enduring works can be given.

There are a variety of ways to test the implications of the linear factor pricing model, both in the time series and in the cross section. For example, in a time-series regression, the CAPM, assuming risk-free lending and borrowing, implies that the regression intercept $\alpha_i$ in Equation 24 should be zero for all test assets.

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i(r_{m,t} - r_{f,t}) + e_{i,t}. \tag{24}$$

Furthermore, a cross-sectional prediction is such that when the average returns on the test assets are regressed against their estimated market betas, the intercept $\lambda_0$ should equal the average risk-free rate, and the slope $\lambda_1$ should equal the expected market risk premium.

$$\bar{r}_i = \lambda_0 + \lambda_i\beta_i + e_i. \tag{25}$$

Fischer Black, Michael Jensen, and Myron Scholes (1972) and Eugene Fama and James MacBeth (1973) are among the earliest testers of the CAPM. Already their tests were able to find some empirical weaknesses in the CAPM. For example, the intercepts $\alpha_i$ in time-series tests were too high (positive) for portfolios having low market betas and too low (negative) for portfolios having high market betas. These problems were also reflected in the cross-sectional tests: The intercept $\lambda_0$ was found to be higher than the average risk-free rate and the slope $\lambda_1$ lower than the average market risk premium. Nevertheless, in spite of these problems, the tests found evidence supporting the main hypothesis of the CAPM that market betas are linearly and positively related to average returns.

Strong empirical evidence against the CAPM started to build up in the 1980s. Rolf Banz (1981) documents that when firm size is measured by market capitalization, investments in a portfolio consisting of stocks of small firms tended on average to yield much higher returns than can be explained by these firms' market betas. This became known as the *small-firm anomaly.* Furthermore, empirical tests have also shown that when firms are assigned to portfolios based on their book-to-market equity (B/M) ratios, portfolios containing high B/M stocks (called *value stocks*) have yielded much higher returns than portfolios with low B/M stocks (called *growth stocks*), and these differences in returns cannot be explained by the portfolios' market betas. This is the value anomaly. In a highly influential paper, Fama and Kenneth French (1992) show that these two variables, size and B/M, characterize the cross section of average stock returns. Even more important, they show that the relationship between market beta and average return is flat. The results imply that the CAPM is unable to explain the average returns in postwar stock market data.

Fama and French (1993) develop a multifactor model that adds two additional factors to the market factor. The first factor is the return difference between a portfolio consisting of high B/M stocks and a portfolio consisting of low B/M stocks. This factor is often denoted high-minus-low (HML). The second factor is the return difference between a portfolio consisting of firms with a small market capitalization and a portfolio consisting of firms with a large market capitalization (small-minus-big, SMB). In the time series, the empirically testable form is

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_{i,MKT}(r_{m,t} - r_{f,t}) + \beta_{i,SMB} SMB_t + \beta_{i,HML} HML_t + e_{i,t}. \quad (26)$$

The model is known as the Fama-French three-factor model. Fama and French (1993) show that the model gives a good description of the average returns on 25 test portfolios that are created by sorting stocks into portfolios based on their B/M ratios and market capitalization. Thus, the model captures the value and size anomalies in the data. Fama and French (1996) show further that the model can also capture some other anomalies in average stock returns. However, the model fails in one aspect: It is unable to explain the short-term momentum in stock returns that was first documented by Narasimhan Jegadeesh and Sheridan Titman (1993). The momentum anomaly refers to the empirical finding that stocks with high past returns (during the previous 3–12 months) continue to offer higher returns during the next year than stocks with low past returns. The momentum anomaly remains as one of the biggest challenges for asset pricing models that are derived under the assumption of rational investor behavior.

The promising empirical performance of the Fama-French three-factor model has made it a benchmark model, against which new asset pricing models developed in the literature are compared. However, the model is not without its criticism. Recall that the factor SMB is the return difference on small and large firms, and HML is the return difference on value and growth firms. Thus, the factors are closely connected to the anomalies they were constructed to explain. One is curious to know what, if any, systematic risk sources these factors are acting as proxy for. There is now a large body of literature that attempts to explain why the betas relative to the Fama-French factors are priced in the cross section of stock returns, but so far, no clear consensus has emerged.

Fama and French (1996) argue that their model can be understood as a form of Merton's (1973) ICAPM. However, because the identity and number of state variables characterizing the evolution of the investment opportunity set are undefined in the theoretical derivation of the ICAPM, it is difficult to evaluate this claim. This ambiguousness involved in choosing the factors for a multifactor representation of asset pricing leads to the danger of data snooping, where researchers, having searched over different data sets, identify variables that appear to be related to the cross section of stock returns in one particular sample. However, there is no guarantee that the variable will work in other time periods or using other test assets. This is a point emphasized strongly by Fama (1991).

Campbell (1996) emphasizes that for tests of the ICAPM, researchers should look for factors that actually forecast the conditional distribution of asset returns, rather than simply choosing any factors that appear to capture common variation in returns. In his empirical tests of an intertemporal asset pricing model, Campbell uses variables that have some forecasting power for future stock market returns. He finds that these variables indeed appear to be priced in the cross section of stock returns.

We saw earlier that the consumption-based CAPM overcomes the problem with selecting factors because only one variable, some function of consumption growth, is enough to theoretically explain asset returns. Unfortunately, so far, the CCAPM has faced problems in explaining the properties observed in real asset return data. Lars Hansen and Kenneth Singleton (1982) assume that investors are described by a power utility function that they are trying to maximize. They then set out to test Equation 23 directly using aggregate consumption data and applying a statistical

method known as the generalized method of moments. Their results imply that the CCAPM cannot simultaneously explain the returns on a stock market index and the risk-free rate. Douglas Breeden, Michael Gibbons, and Robert Litzenberger (1989) find that consumption betas from the CCAPM show a performance comparable to the market betas from the CAPM in explaining stock returns.

What is even more daunting for the consumption-based framework is that in its simplest form, it gives wildly counterfactual predictions about the quantities that it is trying to explain. Rajnish Mehra and Edward Prescott (1985) show that in a simple consumption-based framework, where investors are assumed to have a time-separable power utility function, the predicted difference in stock market returns and the risk-free rate of return is much lower than what is observed in the data. This is the *equity premium puzzle:* The only way to generate the observed equity premium in the simple CCAPM framework is to assume that investors are unreasonably risk averse. However, it must be emphasized that this problem is not unique to the CCAPM. Because most asset pricing models can be viewed as specializations of the basic consumption-based framework, the equity premium puzzle lies hidden in all these models also (see Cochrane, 2005, for a thorough analysis of this issue).

Because of its disappointing performance, the standard version of the CCAPM has been modified in many directions, both in the theoretical and in the empirical research. One of the theoretical modifications has been the development of new utility functions that appear to produce a better match between the theory and empirical findings. On the empirical side, Martin Lettau and Sydney Ludvigson (2001) allow the consumption betas to be time varying in response to different states of nature that they substitute with a well-chosen indicator variable. Their empirical results indicate that value stocks are more highly correlated with consumption growth in bad times, represented by their indicator variable, than in good times, thus leading these stocks to have high expected returns as a compensation for their risks. Furthermore, they also document that these conditional versions of the CCAPM and the CAPM perform almost as well as the Fama-French three-factor model in explaining the returns on the 25 portfolios sorted on book-to-market equity ratios and size that were used as test assets by Fama and French (1993).

## Practical Implications

The previous section described that in the past, value stocks, small stocks, and stocks with high past returns have provided higher returns than their counterparts. Furthermore, these return patterns are left unexplained by the traditional asset pricing models such as the CAPM. Academic research, working under the paradigm that higher returns should arise only as a compensation for higher systematic risk, has mainly interpreted this finding as an indication that the CAPM does not capture all the priced sources of risks in stock returns. However, it is also possible to view these results from an opposite angle. Maybe the CAPM does capture the relevant sources of risk that investors care about. In this case, the returns on value stocks, small stocks, and momentum strategies are not aligned with their riskiness, and consequently, they constitute bargains to investors: By investing in these stocks, investors could historically have achieved high returns without having to take a corresponding high exposure to systematic risk. Perhaps not surprisingly, there are nowadays mutual funds that aim to capitalize on these anomalies and advertize themselves as, for example, following various value strategies.

Thus, if we really believe in the predictions of a specific asset pricing model, then observed deviations from the model can produce valuable investment advice. However, if one is willing to follow such a strategy, it is important to remember some caveats. First, it is possible that the asset pricing model, such as the CAPM, is not well specified and does not capture the priced sources of risk. Second, if such deviations from the risk-return paradigm really do exist, then these deviations are likely to quickly dissipate because of the actions of other market participants who are also trying to capitalize on these anomalies. In line with this argument, the value and size anomalies have weakened considerably in the data since they were made public by academic research. Third, past high returns on a trading strategy do not necessarily imply that the strategy will be profitable in the future. This last point becomes even more significant if the high past returns on a trading strategy cannot easily be explained by an asset pricing model that would indicate that the high returns are compatible with a capital market equilibrium.

## Future Directions for Research

While our understanding of the determinants of risk premiums on various assets has significantly improved since the 1960s, when the CAPM was developed, the picture that the large body of research paints before our eyes is still far from completely clear and coherent. For example, the value anomaly tells us that there is large variation in average returns between value stocks and growth stocks, whereas the momentum anomaly implies that prior winners (stocks with high returns in the past) continue to provide higher returns than prior losers (stocks with past low returns). Asset pricing theory tells us that these high returns should be a compensation for the riskiness of value stocks and prior winners. The asset pricing literature has proposed several new risk factors that aim to explain these return patterns, and these propositions have often received support in the empirical tests that the authors have subjected their models to. However, because many of the proposed models appear to have little in common with each other, each building on its own set of economic arguments, one is still hungry for a coherent

story that would produce a synthesis of the large field of explanations. Furthermore, Jonathan Lewellen, Stefan Nagel, and Jay Shanken (in press) demonstrate that many of these proposed models do not work as well as originally thought when the set of test portfolios is expanded to include returns on portfolios that were not included in the original tests. Ideally, a well-defined asset pricing model should be able to explain the return on any economically interesting portfolio of assets. Thus, much important work remains to be done before we can fully understand the determinants of asset prices.

## Conclusion

This chapter has discussed how financial economists view the relationship between risk and return and how this relationship is reflected in different asset pricing models. The review started with describing the general framework for asset pricing. It was demonstrated that many important asset pricing implications can be derived from an investor's optimal portfolio–consumption choice. An important implication of the analysis was that higher risk should be connected to higher expected returns in market equilibrium and that it is the systematic risk of the asset that matters for pricing purposes. Furthermore, there was an introduction to the important concept of the stochastic discount factor that assigns prices to different assets. Then, there was discussion of some specific asset pricing models and a brief review of their empirical performance. Finally, this chapter offered some practical implications and future directions for research.

## Notes

1. The assumption that there exists a representative agent makes it possible to study the actions of only one agent and use his or her actions to derive some general market-wide asset pricing implications.
2. These securities are also known as Arrow-Debreu securities, contingent claim securities, state securities, and pure securities.
3. Tobin (1958) shows that mean-variance preferences can be motivated if the investors have a quadratic utility function or if the returns are normally distributed.
4. The market portfolio is the sum of all individual holdings of risky assets.

## References and Further Readings


Banz, R. W. (1981). The relationship between return and the market value of common stocks. *Journal of Financial and Quantitative Analysis, 14,* 421–441.

Bernstein, P. L. (2005). *Capital ideas: The improbable origins of modern Wall Street.* New York: John Wiley.

Black, F., Jensen, M., & Scholes, M. (1972). The capital asset pricing model: Some empirical tests. In M. C. Jensen (Ed.), *Studies in the theory of capital markets* (pp. 79–121). New York: Praeger.

Breeden, D. T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics, 7,* 265–296.

Breeden, D., Gibbons, M., & Litzenberger, R. (1989). Empirical tests of the consumption-oriented CAPM. *Journal of Finance, 44,* 231–261.

Campbell, J. Y. (1996). Understanding risk and return. *Journal of Political Economy, 104,* 298–345.

Cochrane, J. (2005). *Asset pricing* (Rev. ed.). Princeton, NJ: Princeton University Press.

Fama, E. F. (1991). Efficient capital markets: 2. *Journal of Finance, 46,* 1575–1617.

Fama, E. F., & MacBeth, J. (1973). Risk, return and equilibrium: Empirical tests. *Journal of Political Economy, 81,* 607–636.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance, 47,* 427–465.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33,* 3–56.

Fama, E. F., & French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *Journal of Finance, 51,* 55–84.

Hansen, L. P., & Singleton, K. J. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica, 50,* 1269–1286.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance, 48,* 65–91.

Lettau, M., & Ludvigson, S. (2001). Resurrecting the (C)CAPM: A cross-sectional test when risk premia are time-varying. *Journal of Political Economy, 109,* 1238–1287.

Lewellen, J., Nagel, S., & Shanken, J. (in press). A skeptical appraisal of asset-pricing tests. *Journal of Financial Economics.*

Lucas, R. E., Jr. (1978). Asset prices in an exchange economy. *Econometrica, 46*(6)*,* 1429–1445.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance, 7,* 77–91.

Mehra, R., & Prescott, E. (1985). The equity premium: A puzzle. *Journal of Monetary Economics, 15,* 145–161.

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica, 41,* 867–887.

Roll, R. W. (1977). A critique of the asset pricing theory's tests. *Journal of Financial Economics, 4,* 129–176.

Rubinstein, M. (1976). The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics and Management Science, 7,* 407–425.

Sharpe, W. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance, 19,* 425–442.

Tobin, J. (1958). Liquidity preference as behavior towards risk. *Review of Economic Studies, 25,* 65–86.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton, NJ: Princeton University Press.

# 21

# Portfolio Theory and Investment Management

Thomas W. Harvey
*Ashland University*

The objective of this chapter is to help readers understand theories of portfolio management. Investment, as opposed to consumption, is the commitment of funds that the investor believes will appreciate in value over time and will provide a return that is sufficient for assuming risk and for exceeding the effects of inflation.

## Portfolio Theory

This chapter begins by reviewing theories of investment and portfolio management that have been prevalent throughout the twentieth century. First, it reviews the firm foundation theories of Benjamin Graham and David Dodd, developed in their seminal book, *Security Analysis,* published in 1934. Following Graham and Dodd was John Burr Williams's (1938) theory of investment value, which added further sophistication to the Graham and Dodd calculation and which became one of the more common contemporary valuation techniques for equity securities. And then, the work of Harry Markowitz (1952) on expected utility theory, modern portfolio theory, and diversification theory furthered the analytics of Graham and Dodd and the valuation methods of Williams. Together, they became the dominant investment techniques in the twentieth century.

Second, the chapter examines Dow theory, which is the basis of technical analysis, or the reading of stock price charts to derive patterns of price movements (Schannep, 2008). Charting price movements has its roots in eighteenth-century Japan, when Munehisa Homma, a commodity merchant from Sakata, created a futures market for rice and used the charts to determine the ways in which he thought future prices would move. Homma's use of the charts is the subject of Steve Nison's (2001) book, *Japanese Candlestick Charting Techniques,* which is considered "the bible of candle charting" and contributed to the increased popularity of chart use in the latter part of the twentieth century.

The third theoretical approach is the random walk, as described by John Maynard Keynes in 1936, refined by Eugene Fama in 1965, and made popular by Burton Malkiel (1973/2003) in his landmark book, *A Random Walk Down Wall Street.* Keynes, Fama, and Malkiel postulated that stock prices were simply functions of investor decisions, and study of financial statements or examination of price charts was a waste of time. Keynes believed that all the investor had to do was find an opportunity before the market discovered it. The assumption was that the market would find it eventually, which would increase the price, creating wealth for those who had purchased it first.

Finally, the emerging field of behavioral finance as presented by Daniel Kahneman and Amos Tversky (1979), Hersh Shefrin (2002), and John Nofsinger (2008), among others, maintains that psychology should be brought into the dialogue about investments. Whether investors believe in fundamental analysis, technical analysis, or the theory of the random walk, they react to some kind of information as they make their resource allocation decisions. Personal

bias, emotion, experience, and even mood influence the way decisions are made as investors move into and out of the market.

## Firm Foundation Theory and Fundamental Analysis

Traditional finance, which was grounded in the rationality of investors, was significantly influenced by Graham and Dodd's (1934) *Security Analysis.* This work established the principles of fundamental analysis for making resource allocation decisions and became the foundation for investment portfolio management for much of the twentieth century.

Graham and Dodd recognized the difference between speculation, which focuses on when to buy or sell a security, and investment, which allows investors to determine what to buy or sell. They prescribed fundamental analysis, the centerpiece of the firm foundation theory, as a thorough study of the strengths and weaknesses of individual companies, primarily concentrating on a review of past financial results, which can then be used to develop a projection of future performance. Those estimates of future earnings became what Graham and Dodd (1934) called the intrinsic value of the firm. The idea was to find companies whose fundamental, or intrinsic, value was more than the current market price of the stock, which, thus, represented a buying opportunity.

To calculate the intrinsic value of a prospective investment meant studying such factors as past earnings and operating margins and making assumptions about expected growth rates to arrive at a projection of future earnings. Specifically, Graham and Dodd studied companies with high earnings-to-price ratios, low price-to-earnings ratios, high dividend yields, prices below book values, and net current asset valuations. They tried to find earnings trends that would represent the underlying value of the business that could be carried into future years.

That data, along with information about the economy and evaluation of the firm's management, could result in a complete understanding of the past performance and future prospects that, then, became the basis for the investment decision. If those prospects were promising, investors would buy the stock; if those prospects were not promising, they would not commit their resources.

Williams (1938) took the concept of intrinsic value a step farther. He agreed with Graham and Dodd (1934) that asset prices should reflect the basic, or fundamental, value of the firm but used different factors in his calculations. Whereas Graham and Dodd focused on historical earnings per share that they projected into the future, Williams devised the discounted cash flow model in order to determine the present value of those earnings as represented by dividends. The intrinsic value, then, was the present value of the firm's future dividends. In 1938, when Williams wrote *The Theory of Investment Value,* most if not all publicly traded companies paid dividends, which made the model most useful. In the latter part of the twentieth century, many firms omitted dividends, preferring to reinvest their earnings for future growth, thus limiting the use of the Williams model. However, the model was modified in the late 1990s to discount future free cash flow instead of dividends to derive the intrinsic value for any firm, which makes it applicable to all firms (Gray, Cusatis, & Woolridge, 1999). The theories of Graham and Dodd (1934) and Williams (1938) suggest that if the market price is less than the intrinsic value, investors will bid the price up to meet that value. Conversely, if the intrinsic value is lower than the market price, investors will bid it down.

As the firm foundation theory and fundamental analysis gained prominence in the middle of the twentieth century, Markowitz (1952) published "Portfolio Selection" in the *Journal of Finance.* In this work, Markowitz presented the mathematics of diversification that had been introduced in 1738 by Daniel Bernoulli. Bernoulli described the nature of the risk associated with a collection of assets (Rubenstein, 2002). Markowitz realized that Williams did not address the issue of risk inherent in assets because Williams believed in the mitigating factor of diversification. That is, holding a variety of assets that were not correlated with each other would eliminate the risk entirely. Further, Markowitz saw that the Williams model was problematic because the future dividend of the firm could not be known for certain. The critical factor, he understood, was the expected return of the investment (Fabozzi, Gupta, & Markowitz, 2002; Varian, 1993).

Thus, Markowitz suggested that investors should construct their portfolios based on the trade-off between risk and reward. This view is grounded in expected utility theory (EUT), which reflects the concept of *homo economicus,* or the economic man who makes financial decisions completely rationally, using all of the information that is available at the time. This is known as modern portfolio theory, and Markowitz suggested that the more risk that was taken, the greater the return the investor should require. That is, investors should try to determine the potential return for their portfolios given the amount of risk that they are willing to assume. The expected return (reward) and the volatility (risk) of a portfolio could be calculated, allowing investors to make more educated decisions about the allocation of their assets (Fabozzi et al., 2002; Graham, 1973; Rubenstein, 2002). A fundamental component of behavioral finance, however, is that investors can have different perceptions of the risk of a potential investment as well as of its expected value.

Markowitz explained that different classes of assets have different amounts of risk associated with them. The theory suggested, for example, that small company stocks would have higher average annual returns than large company stocks, corporate bonds, and government bonds, but they also would have the greatest amount of overall risk. The history supports the theory, as shown in Figure 21.1. Small companies are higher risk, but they also have higher returns. In sum, Markowitz argued that a balance of risk

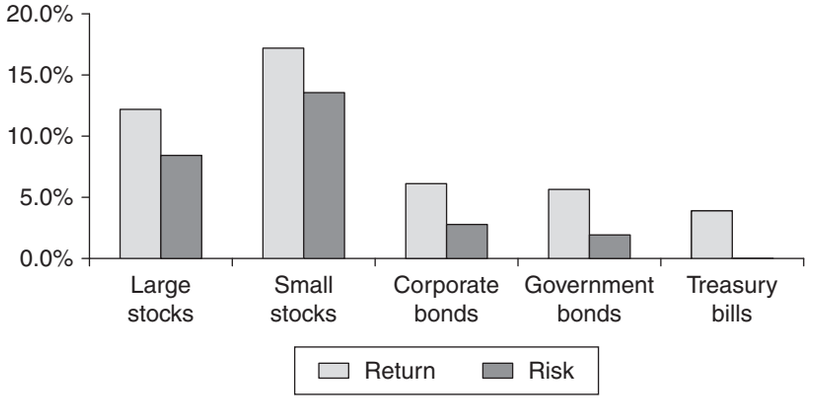


**Figure 21.1** Average Annual Returns and Risk Premiums: 1925–1996

SOURCE: Jordan and Miller (2008, p. 15).

and reward could be struck according to investors' expectations for reward, balanced against their tolerance for risk. Investors could reduce the risks associated with individual stocks, depending on the total amount of risk being assumed, by diversifying their portfolios with 20 to 30 stocks (Markowitz, 1952, 1976).

Traditional finance, then, was based on Graham and Dodd's (1934) concept of intrinsic value; Williams's (1938) theory of investment value and the present value calculation; and modern portfolio theory, diversification theory, and expected utility theory of Markowitz (1952).

## Dow Theory and Technical Analysis

In addition to conducting fundamental analysis of the firm, studying the price charts—or technical analysis—is a way to evaluate potential investments as well. Technicians believe that prices are the results of all the factors in the market, including earnings releases, economic data, investor emotions and biases, and especially the result of previous price changes. Fundamental analysis is useful for technicians, but price momentum is the key. While the use of charts was originated by the 1700s by Munehisa Homma in the prediction of the future price of rice (Nison, 2001), Charles Dow laid the groundwork for the use of charts early in the twentieth century. While Dow did not fully document his theory in a book or published article, he did document his ideas about stock price movements in *The Wall Street Journal,* of which he was editor in the last decade of the nineteenth century. Several years later, William Peter Hamilton formalized Dow theory in *The Stock Market Barometer* in 1922, which was followed by the definitive *Dow Theory,* written by Robert Rhea in 1932 (Schannep, 2008).

Dow theory maintained that past price movements could signal future price movements (Schannep, 2008). It centered on the primary trend, secondary patterns, and daily fluctuations and theorized that the primary trend would last for a long period of time while secondary patterns and daily price changes were important only when they impacted the primary trend. Charting the price movements allowed investors and speculators to identify those trends that they believed provided clues as to the next price change. Investors bought and sold according to what they saw.

Stock prices anticipate earnings, and Dow understood that when business is good, the price of the firm's stock would reflect it and vice versa. That is, the market would gauge a firm's future prospects, and if the future were promising, the price would start to advance. A pattern would start to develop as the price continued upward, for example, which would signal a buying opportunity. However, on the other hand, it might also be a signal to sell if the trader thought a reversal is imminent. The decision-making process is one of human judgment.

By looking at the charts, technicians can see the lowest price at which a particular stock has traded over a specified period of time, which is known as the support level (Schwager, 1996, 1999). That tells the analyst that there is a strong likelihood that future prices will not go below it. Conversely, the highest price at which the stock has traded can also be seen and is known as the resistance level. That price is the one that will not, in all likelihood, be exceeded. However, there are no guarantees when dealing with stock prices. Prices can break through the resistance level if the market senses pending improvement in future earnings, or there can be a break through the support level if those future prospects become clouded.

The trend line for a particular stock can be compared to the Dow Jones Industrial Average, NASDAQ (the National Association of Securities and Dealers Automated Quotation), or S&P (Standard and Poor's) 500 to gauge performance over a period of time, whether it be 5 days, 5 years, or somewhere in between. That comparison shows whether the stock has done better or worse than the index, which will have an influence on the investor's decision. Comparisons can also be made to competitors so that the analyst can

evaluate the strength of the trends of the competitive prices. Analysts can also compare the price trend to moving averages of its performance (50 day or 200 day), its relative strength index (RSI), or the volume of shares being traded as they try to discern the direction in which prices will move.

There are many different technical analytics that technicians use, such as bar charts, candlestick charts, point and figure charts, head and shoulders diagrams, saucers, double tops, double bottoms, and fulcrums (Schwager, 1996, 1999). Because there are so many applications and uses of technical analysis, most technicians have a select few methods that they use. Just as with fundamental analysis, technical analysis requires great discipline. Historical patterns and momentum can be useful to determine future direction; however, the future does not always correspond directly with the past.

## The Random Walk and the Efficient Market Hypothesis

Eugene Fama (1965) instructed that "stock price changes have no memory—the past history of the series cannot be used to predict the future in any meaningful way. The future path of the price level of a security is no more predictable than the path of a series of cumulated random numbers" (p. 56). Thus, the theory of the random walk is another way to look at asset pricing.

Fama (1965) wrote that the "techniques of the chartist have always been surrounded by a certain degree of mysticism" (p. 55). Thus, analysts, Fama argued, would turn to fundamental analysis and the concept of intrinsic value. But he conceded that there could be different estimations of intrinsic value that would affect actual prices, a concept that is at the foundation of behavioral finance. Price movements, then, could be considered random.

Malkiel (1973/2003) and other proponents of the theory of the random walk believe that there is no relationship between past and future price movements. That is, future prices cannot be predicted. "Random walkers claim that the stock market adjusts so quickly and perfectly to new information that amateurs buying at current prices can do just as well as the pros" (p. 175). That suggests that fundamental analysis and technical analysis are of no value in the attempt toward investment success.

The theory of the random walk is related to the castle-in-the-air theory, which was originated by the British economist John Maynard Keynes (1936/1965) in the 1930s and simply emphasized analyzing the behavior of the market. Under this theory, investment decisions should be made in response to what the crowd is going to do. Keynes maintained that if he could identify an investment opportunity before the rest of the market did, he could take advantage of that opportunity when other investors began to see it as well. Castles in the air are representations of "herd mentality," wherein investors purchase assets because they see others doing the same and do not want to be left out. As such, prices continue to increase when there may be no rational reason for it (Malkiel, 1973/2003).

It should be noted that many investors have done quite well building castles in the air. The tulip craze in Holland, the South Sea bubble, the dot-com bubble, and most recently, the housing run in the first decade of this century were such instances where crowd watchers earned significant returns. But eventually, all bubbles burst. The ideal strategy is to buy early, enjoy the price appreciation, and then sell before the bubble bursts because history has shown that after each major market craze, prices almost always stabilize between their normal support and resistance levels.

Malkiel also presented the concept of market efficiency, in the form of the efficient market hypothesis (EMH). EMH proposes that current market prices incorporate and reflect all relevant information about a company or a stock as soon as that information is known. The theory maintains that stocks always trade at a fair value, so fundamental and technical analyses are useless for finding undervalued equities or predicting market trends. It does not claim that a stock's price reflects future financial performance; rather, the stock price is a function of investor reaction to information that has come into the market.

Malkiel (1973/2003) wrote, "Efficient-market theory . . . [is] built on the premise that stock-market investors are rational" (p. 216). However, the recent development of behavioral finance has produced a strong argument for investor irrationality. Malkiel acknowledged that many investors were influenced by psychological biases and thus may make irrational decisions, but he also believed that investors are balanced by rational market participants. In the long term, the market is expected to continue upward. However, short term, market behavior is completely random. "Even if investors are irrational in a similar way, efficient-market theory believers assert that smart rational traders will correct any mispricings that might arise from the presence of irrational traders" (p. 217).

## Behavioral Finance

No one who has ever experienced or observed the capital markets can deny the evidence that psychology is a prominent issue in the pricing of securities. Shefrin (2002) maintained, "Psychology is hard to escape; it touches every corner of the financial landscape, and it's important" (p. 309). The study of the effects that psychological factors have on investment decisions and thus on the financial markets has been termed *behavioral finance.*

Those who ascribe to the tenets of behavioral finance do not reject the foundations of traditional finance, which holds to the theories of Graham and Dodd (1934) and Williams (1938). Further, they do not eschew technical analysis (Nison, 2001; Schannep, 2008; Schwager, 1996, 1999) or even the castles-in-the-air theory of Keynes (1936/1965). However, the field of behavioral finance

does maintain that psychological influences cause prices to deviate from their fundamental values (Shefrin, 2002).

Robert Olsen (1998) suggested that there is increased interest in behavioral finance because of empirical studies that show that traditional theories of finance have become deficient and because of the work of Kahneman and Tversky (1979) in the area of prospect theory. Kahneman and Tversky offered prospect theory as an alternative to expected utility theory as the field tried to understand, and then to explain, the ways in which investors make decisions, as opposed to the way in which the purest view of traditional finance sought to explain it. Traditional finance was, indeed, based on the concept of *homo economicus,* who in Olsen's (1998) view, was a decision maker who had unlimited amounts of information and who was able to process all of that information logically and rationally in the attempt to maximize utility.

Meir Statman (1995) provided a relatively simple distinction between traditional finance and behavioral finance. In the view of behaviorists, traditional finance is based on how human beings should behave, and behavioral finance focuses on how people actually do behave (Baker & Nofsinger, 2002; Barber & Odean, 1999; Curtis, 2004). If, in fact, investors were rational, the bubbles (like the tulip craze in Holland in the 1600s, the escalation of the NASDAQ in the late 1990s, and the housing run-up of the early part of the first decade of the twenty-first century) would not have occurred. The panic of 2008 would not have happened either. However, as Louis Lowenstein (2006) pointed out, investors can strive to be as rational as possible.

Behavioral finance is based on the use of heuristics, which Hubert Fromlet (2001) defined as the use of experience and practical efforts to answer questions or to improve performance. But that raises the question of the differences in the experiences of individual investors and their psychological reactions to those experiences that shape and influence their thinking about opportunities that might be available in the market. In sequence, then, two investors with different experiences and reactions to those experiences might view the same opportunity totally differently because of their biased views. Edgar Peters (2003) instructed that the use of heuristics might outweigh the use of statistics because the speed of investment management has increased because of the Internet.

## Investment Management

Management of an investment portfolio is accomplished by first establishing the investor's objectives.

### Objectives

Understanding the difference between speculation and investment is critical to the way in which potential alternatives for resource allocation are evaluated. Speculation occurs in the very near term and is based on price movements as traders try to take advantage of even the narrowest of spreads between the bid and the ask prices. The analysis is intuitive because the speculators see a price movement and act on it. There is little time, usually, for more thorough analysis in the evaluation process. Investment, on the other hand, is for the long term where price appreciation and dividends matter.

### Investment Decision

Investors adopt a range of strategies, depending on their objectives. Investments can be made in government securities, bearing minimal, if any, risk of default and earning a low rate of interest. Investors could also put resources in certificates of deposit in the local bank, which will be insured by the Federal Deposit Insurance Corporation (FDIC) and which will earn a low rate of interest as well.

A second strategy is to build a diversified portfolio of stocks, bonds, and cash. Depending on the investor's risk tolerance, this portfolio could be weighted more toward bonds than stocks or vice versa. The investor would also determine whether the strategy would be to have a growth portfolio, where price appreciation is the goal, or a value portfolio, which stresses the importance of dividends. On the other hand, Warren Buffett advises that investors should seek out a select group of well-managed companies instead of diversifying as prescribed by Markowitz (1952).

Bonds represent a claim on the issuer's assets and are a contract with the issuer, as opposed to stocks, which represent ownership that is not contractual. The bond contract, called the indenture, states the terms and conditions of the issue, especially the timing and amounts of the interest payments and principal repayments. Generally, for most corporate and government bonds, interest is paid semiannually.

The first step in creating a bond position is to determine the type of bond to be included in the portfolio. Bonds are issued by the federal government, agencies of the federal government, state and local governments, foreign governments, and corporations. Government bonds are backed by the full faith and credit of the United States, which means that there is very little chance of default. Those issued by states and municipalities are supported by the taxing power of the governmental unit issuing them, which also means little or no risk. Or these entities can issue revenue bonds, which are supported by the income that is generated by specific projects. Another option is the acquisition of bonds of foreign governments, which have considerably more risk associated with them.

Corporate bonds are backed by the earning power of the company issuing them, and the decision to include them in the investor's portfolio is generally based on the credit quality of the issue. In most cases involving corporate

bonds, investors choose a credit rating that is at least investment grade, as shown in Table 21.1. The rating of the bond is not the agency's assessment of the overall quality of the company issuing the bonds. Rather, it is the evaluation of the particular bond issue as supported by the company's financial strength, the quality of its management, and its current position in its markets. Bonds with the highest ratings represent the highest creditworthiness of the issuer and will have the lowest interest rate because they have the lowest risk.

Other characteristics of a bond are the coupon rate (the rate of interest that will be paid over the life of the bond); the maturity date when the bond will be redeemed; the call provision, which allows the issuer to redeem the bond prior to maturity; and any conversion provision, which enables the issuer to convert the bond into shares

**Table 21.1** Rating of Corporate Bonds

| *Rating Agency* | | | |
|---|---|---|---|
| *Moody's* | *Duff & Phelps* | *Standard & Poor's* | *Credit Rating Description* |
| | | | **Investment-Grade Bond Ratings** |
| AAA | 1 | AAA | Highest credit rating, maximum safety |
| Aa1 | 2 | AA+ | |
| Aa2 | 3 | AA | |
| Aa3 | 4 | AA– | |
| A1 | 5 | A+ | |
| A2 | 6 | A | Upper medium quality, investment-grade bonds |
| A3 | 7 | A– | |
| Baa1 | 8 | BBB+ | |
| Baa2 | 9 | BBB | Lower medium quality, investment-grade bonds |
| Baa3 | 10 | BBB– | |
| | | | **Speculative-Grade Bond Ratings** |
| Ba1 | 11 | BB+ | Low credit quality, speculative-grade bonds |
| Ba2 | 12 | BB | |
| Ba3 | 13 | BB– | |
| B1 | 14 | B+ | Very low credit rating, speculative-grade bonds |
| B2 | 15 | B | |
| B3 | 16 | B– | |
| | | | **Extremely Speculative Bond Ratings** |
| Caa | 17 | CCC+ | |
| | | CCC | |
| Ca | | CCC– | |
| C | | CC | Extremely speculative |
| | | C | |
| | | D | Bonds in default |

SOURCE: Jordan and Miller (2008, p. 607).

of common stock. Corporate bonds have varying maturity dates and coupon rates and typically have a face value of $1,000.

As with any investment, the potential return can be calculated and then compared against competing assets. The formula for calculating the present value of a bond is

$$V = \sum_{t\ 1}^{n} \frac{CT}{(1+i)^{t}} + \frac{Pn}{(1+i)^{n}},$$

where

*V* = the market value or price of the bond
*n* = the number of periods
*t* = each period
*Ct* = Coupon rate or interest rate for each period
*Pn* = Par value or maturity value
*i* = interest rate in the market

The price of any bond represents the present value of the interest payments over the maturity of the bond and the present value of the par value when the bond matures.

If a bond has a par value of $1,000, pays 6.00% interest for 15 years, and the current interest rate in the market is 8%, what is the value of the bond?

Using present value tables or a financial calculator,

n = 30 (interest on bonds is paid semi-annually)

Pn = $1,000

i = 4%

C = 3.00% (interest is paid semi-annually) or $30

Present value of the bond

$827.08

The first step in the calculation is to compute the interest that the bond will pay. Because interest is paid semiannually, divide the annual rate (6%) by 2. The result is 3%. Taking 3% of the par value of $1,000 yields interest payments of $30 every 6 months. Similarly, the current interest rate in the market is 8%, but that also has to be halved because the calculation is really dealing with 6-month intervals.

So the bond is worth $827.08 today and will provide the investor a steady income stream, but what is the current yield that will allow for comparisons of competing investments to be made?

$$CY = \frac{C}{V}$$

$$CY = \frac{\$30}{\$827.08}$$

$$CY = 3.63\%$$

In this circumstance, the current yield is 3.63%, which allows the investor to compare it to competing alternatives. Using this analysis, developing a bond portion of the portfolio can assist in reducing the amount of risk that the investor assumes in consideration with the objectives that have been set.

Mutual funds are another possible investment for the portfolio, and they, too, must coincide with the investor's objectives. The Investment Company Institute (ICI, 2008) categorized mutual funds as the following:

- Stock funds
- Hybrid funds
- Taxable bond funds
- Municipal bond funds
- Taxable money market funds
- Nontaxable money market funds

In addition, ICI reported that, as of September 2008, these funds held $10.631 trillion of investor assets.

Mutual funds may be open-ended, which means that shares of the fund can be traded at any time. Or funds may be closed-ended, which means that the number of shares in the fund is fixed and can be traded only if the investor can find another investor who wishes to engage in a transaction, either buying into the fund or selling out of it.

In addition to incurring management fees, investors in mutual funds may incur fees for purchasing the fund, called loads; reimbursing the fund for marketing expenses, called 12B-1 fees; and trading within the fund.

The investor can also purchase index funds, which as the name implies, are funds that mirror an index like the S&P 500 or the Wilshire 5000. These funds rise and fall as the index moves up and down and have the same amount of risk as the index. Exchange-traded funds (ETFs) are also available. These are funds that resemble closed-end mutual funds and index funds but can be more specialized. That is, there are numerous ETFs that consist of stocks of companies in a particular country, industry, sector, or commodity.

Another investment alternative is to purchase preferred stock in some of the larger companies being traded on the New York Stock Exchange and on NASDAQ. Preferred stock is a hybrid between common stocks and bonds, with the dividend being senior to the common dividend, but the claim on the company's assets is subordinate to the bondholders and other creditors. Preferred stock represents ownership in the company, but it is nonvoting, which means that the preferred shareholder has no say in the management of the organization, unlike common stockholders.

The final investment alternative in this discussion is common stock, which is traded around the world on various exchanges every day, most notably in the United States on the New York Stock Exchange and on NASDAQ. Common stock represents voting ownership, which means that the common shareholder is consulted

on matters affecting the corporation as a whole, such as merger and acquisition activity, the membership of the board of directors, or the selection of the independent auditing firm.

As was shown in Figure 21.1, common stock represents the most risk of any of the alternatives that have been presented, with the exception of derivatives, and stockholders demand a higher return for taking that risk, called the equity risk premium. The greater the risk of a common stock, the higher the equity risk premium demanded by the investor.

The formula for calculating the return on common stock is as follows:

$$\text{Percentage return} = \text{Capital yield gain} + \text{Dividend yield gain}$$

$$\text{Percentage return} = (Pt + 1 - Pt) / Pt + Dt + 1 / Pt,$$

where
$Pt$ = stock price at the beginning of the year
$Pt + 1$ = stock price at the end of the year
$Dt + 1$ = annual dividend paid on the stock

The purchase price of ABC Corporation was $90 per share. At May 31, the stock was selling for $106 per share. ABC paid a quarterly dividend of $0.75. If the investor had purchased 1,000 shares at $90 per share, what was the total percentage return?

R = (Pt + 1 – Pt) / Pt + Dt + 1/ Pt

R = ($106 - $90) / $90 + $3 / $90

R = $16 / $90 + $3 / $90

R = .1778 + .0333

R = .2111

R = 21.11%

Calculating the historical and expected return allows the investor to compare the stock's performance with the same returns on other investment alternatives so that a more educated decision can be made.

To understand the nature of the volatility involved with securities, there are two types of risk associated with every share of common stock. First, there is the risk of the market, which is called systematic risk. This is the risk that every stock faces, such as interest rate increases and decreases or changes in the money supply. And then, there is unsystematic risk, which is that associated with a specific company. As Markowitz (1952) noted, by diversifying the portfolio with 20 to 30 carefully chosen stocks, that risk can be mitigated and even eliminated if the movements in the stock prices that are in the portfolio are not correlated.

Systematic risk is measured with beta. Stocks with a beta of 1 have just as much risk as the market. Those with a beta of greater than 1 are riskier than the market. That is, if the beta is 2, that stock is twice as risky as the market. Similarly, if the beta is 0.5, the stock is half as risky as the market. Portfolio beta is measured by the weighted-average beta of the entire portfolio, as follows:

$$\beta p = Xa\beta a + Xb\beta b + Xc\beta c + \ldots Xn\beta n,$$

where
$\beta p$ = portfolio beta
$X$ = percentage of assets allocated to each stock
$\beta$ = beta associated with each stock

To measure the actual return on the portfolio, we use Jensen's alpha, which is the return on the portfolio less the expected portfolio return, as follows:

$$\alpha = Rp - \{Rf - [E(Rm) - Rf] \times \beta p\},$$

where
$Rp$ = actual return on the portfolio
$Rf$ = risk-free rate
$E(Rm)$ = expected return on the market
$\beta p$ = weighted average beta of the portfolio

Before the alpha can be computed, a portfolio is constructed according the investor's objectives. If the investor is a risk taker, the portfolio would be directed toward growth and would have a fairly high beta. On the other hand, if the investor is risk averse, a value portfolio would be constructed with beta being 1 or less.

## Recent Changes in Market Behavior

The subject of portfolio theory and investment analysis has taken on new meaning as the United States endured the financial crisis of the latter portion of the first decade of the twenty-first century. In 2007 through 2009, the country experienced a financial collapse that saw the demise of the investment banking industry, unprecedented housing foreclosures, significant job losses, and the drastic decline in the stock market.

The collapse of the nation's financial sector and the unprecedented sell-off in equity markets was partly a function of the speed with which financial information arrived and the creation of sophisticated derivative securities, which added significant volatility to the market (Lewis, 2004). Trading volume soared as the result of computerized institutional trading systems and online trading systems for individuals. The average volume of

stocks listed on the Dow Jones Industrial Average in September and early October 2008 was 7.2 billion shares per day, almost double the amount from the previous January. Not only was there more stock being traded, but the holding period for stocks decreased as well, further adding to volatility. Internet-based financial systems enable individuals to trade quite easily, but 57% of the respondents to a recent survey reported that they conducted no transactions using the Internet during the previous 12 months (ICI, 2008). That result suggests that it is the institutions, such as pension funds, trust companies, banks, and hedge funds, that are driving the volume volatility and potentially the price.

## Conclusion

The purpose of this chapter has been to trace the development of dominant theories of investment management. Investment theory has moved from traditional finance and its emphasis on *homo economicus* to behavioral finance, which brings elements of psychology and emotion into the human decision making associated with balancing risks and returns in financial markets.

Behavioral finance recognizes that biases, emotions, backgrounds, and experiences influence investor decision making. These human traits may skew the ability to make rational and logical resource allocation decisions as theorized in traditional finance. Thus, there is room for the elements of rational study and examination of various investment alternatives prior to making those decisions.

## References and Further Readings


Baker, H., & Nofsinger, J. (2002). Psychological biases of investors. *Financial Services Review, 11,* 97–116.

Barber, B., & Odean, T. (1999, November–December). The courage of misguided investors. *Financial Analyst Journal,* pp. 41–55.

Curtis, G. (2004, Fall). Modern portfolio theory and behavioral finance. *Journal of Wealth Management,* pp. 16–22.

Fabozzi, F., Gupta, F., & Markowitz, H. (2002). The legacy of modern portfolio theory. *Journal of Investing, 11*(3), 7–22.

Fama, E. (1965). Random walks in stock market prices. *Financial Analysts Journal, 21*(5), 55–59.

Fromlet, H. (2001, July). Behavioral finance—theory and practical application. *Business Economics,* pp. 63–69.

Graham, B. (1973). *The intelligent investor: A book of practical counsel* (4th ed.). New York: HarperBusiness.

Graham, B., & Dodd, D. (1934). *Security analysis.* New York: McGraw-Hill.

Gray, G., Cusatis, P., & Woolridge, R. (1999). *Streetsmart guide to valuing a stock: The savvy investor's key to beating the market.* New York: McGraw-Hill.

Investment Company Institute, Securities Industry, & Financial Markets Association. (2008). *Equity and bond ownership in America.* Washington, DC: Authors.

Jordan, B., & Miller, T., Jr. (2008). *Fundamentals of investments: Valuation and management* (4th ed.). New York: McGraw-Hill Irwin.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291.

Keynes, J. M. (1965). *The general theory of employment, interest, and money.* New York: Harcourt Brace. (Original work published 1936)

Lee, H.-J., Park, J., Lee, J.-Y., & Wyer, R. S., Jr. (2008). Disposition effect and underlying mechanisms in e-trading of stocks. *Journal of Marketing Research, 45,* 362–378.

Lewis, M. (2004). *Moneyball.* New York: W. W. Norton.

Lowenstein, L. (2006). Searching for rational investors in the perfect storm: A behavioral perspective. *Journal of Behavioral Finance, 7*(2), 66–74.

Malkiel, B. (2003). *A random walk down Wall Street.* New York: W. W. Norton. (Original work published 1973)

Markowitz, H. (1952). Portfolio selection. *Journal of Finance, 7*(1), 77–91.

Markowtiz, H. (1976). Investment for the long run: New evidence for an old rule. *Journal of Finance, 31*(5), 1273–1286.

Nison, S. (2001). *Japanese candlestick charting techniques* (2nd ed.). New York: New York Institute of Finance.

Nofsinger, J. (2008). *The psychology of investing* (3rd ed.). Upper Saddle River, NJ: Pearson Education.

Olsen, R. A. (1998, March–April). Behavioral finance and its implication for stock price volatility. *Financial Analysts Journal,* pp. 10–18.

Peters, E. (2003). Simple and complex market efficiencies: Integrating efficient markets, behavioral finance, and complexity. *Journal of Behavioral Finance, 4*(4), 225–233.

Rubenstein, M. (2002). Markowitz's "portfolio selection": A fifty-year retrospective. *Journal of Finance, 55*(3), 1041–1045.

Schannep, J. (2008). *Dow theory for the 21st century.* New York: John Wiley.

Schwager, J. (1996). *Technical analysis.* New York: John Wiley.

Schwager, J. (1999). *Getting started in technical analysis.* New York: John Wiley.

Shefrin, H. (2002). *Beyond greed and fear.* New York: Oxford University Press.

Statman, M. (1995). A behavioral framework for dollar-cost averaging. *Journal of Portfolio Management, 23*(2), 8–15.

Varian, H. (1993). A portfolio of Nobel laureates: Markowitz, Miller, and Sharpe. *Journal of Economic Perspectives, 7*(1), 159–169.

Williams, J. B. (1938). *The theory of investment value.* Cambridge, MA: Harvard University Press.

# PART III

# PUBLIC ECONOMICS

# 22

# Externalities and Property Rights

Mahadev Ganapati Bhat
*Florida International University*

The private market is a common economic mechanism for allocating scarce economic resources in an efficient manner. Imagine a situation where all decisions (when, where, how much, and at what price) relating to producing and consuming every good or service in the society are made by government or a group of individuals. Under such a system, one can expect much wastage, time delay, and lack of coordination among decision entities, which would hamper the smooth and efficient delivery of goods and services. In an ideal world, perfectly competitive private markets are means to achieving the most efficient way of making goods and services available to people, of exactly the quantities they need, at prices they can afford, and at a time and place they need.

However, it is rare to find such a perfect competitive market. Many conditions are necessary for the smooth functioning of the private market (McMillan, 2002). They are as follows: (a) Side effects of production and consumption are limited, or there are no *externalities* in production or consumption; (b) there are well-defined *property rights* to resources, products, and production technology; (c) there exists widely available information about buyers, sellers, and products; (d) enforceable private contracts are present; and (e) the conditions of competition exist. More often than not, one or more of the above conditions are violated in private markets, leading to suboptimal market performances, in terms of product quantity, quality, and prices and consumers' and producers' satisfaction. In this chapter, we focus on the first two assumptions of markets: absence of externalities and well-defined property rights. We will first narrate these two concepts and their nuances and then discuss how these concepts are related to one another. Later, we will present the adverse effects on the market system of violating the above two assumptions, as well as regulatory and market means to correct these adverse effects.

## Theory

### Externalities

Many economic decisions that people make in their day-to-day lives have economic consequences beyond themselves. We can find examples wherein certain people may bear costs or gain benefits from the actions or decisions taken by others, even though they had no role in making those decisions. A profit-maximizing entrepreneur normally takes into account only those out-of-pocket costs that she pays for (i.e., the costs of inputs used in the production process) while deciding how much, how, what, and when to produce goods and services. Such costs are called private costs. To increase profits, the entrepreneur always looks for ways and means—via cheaper technology and inputs of production—to keep the production costs low. It makes no economic sense for her to consider any likely costs of this economic decision inflicted on people or entities other than herself or her firm.

For instance, a farmer has to plow his land to maintain proper tilth of the soil and apply chemicals to kill weeds and pests. Excessive plowing of land with poor soil condition causes soil erosion from his land and contaminates downstream water bodies with soil sediments. The deposition of soil sediment may decrease the water storage capacity of the downstream lake and suffocate fish and other aquatic creatures. Residues from chemical application on the farmland may also end up in the lake, making that water body unsuitable for fish. Both of the above

actions on the part of the farmer might mean increased costs of lake water storage and loss of fish production. In the absence of any legal, social, or moral obligation, the farmer upstream has no incentive to reduce soil erosion or chemical application as long those actions bring him economic profits. Nor does the farmer feel responsible for paying for those costs or income losses to downstream water users. Such costs or negative income impacts are called *negative externalities* or *external costs.*

Consider another example. There is a private company selling cell phones to customers. The amount of profit the company makes from this product depends on the price and the number of phones sold in the market. The buyers are looking forward to buying a phone at as cheap a price as possible. The price that any given buyer is willing to pay equals the marginal value that he derives from the phone. This marginal value accounts for only his personal (private) value that he derives from it. Note that there could be other buyers who may have already bought this product. When this buyer makes the purchase, he becomes the newest member of the existing network of cell phone users. Each new user enhances the utility or value of the phones owned by all the existing customers. The larger the number of users of a given product is, the greater the ability of all users is to stay in touch with each others. As in the case of negative externality, this private buyer would consider only his private benefits when he makes his purchasing decision. There is no incentive for him to worry about what others might gain from his decision. The other network members are no part of his decision. The benefits that he brings to all other network members are an example of *positive externality* or *external benefit.*

In both these examples, when a particular economic decision imposes economic costs on, or generates economic benefits to, a third party, the behavior of the decision maker in the market does not include the costs to, or the preferences of, those who are affected. This means that the *private costs* borne or *private benefits* enjoyed by a decision maker are much different (normally lower) from the *true costs* or *true benefits* of that decision. The true costs and true benefits are also called *social costs* and *social benefits,* respectively. The difference between the social costs and private costs are negative externalities (external costs), and the difference between the social benefits and private benefits are positive externalities (external benefits).

### *Definition of Externality*

Now we are ready to define externality in a more precise fashion. According to Baumol and Oates (1988), an externality is an unintentional effect of an economic decision made by persons, corporations, or governments on the consumption or production by an outside party (person, corporation, or government) who is not part of the original decision. It is evident from this definition that externalities can be present in both production and consumption activities. Also, for an externality to exist, there needs to be a direct connection between the parties through real (nonmonetary) variables. Let us revisit the earlier examples of negative and positive externalities. For instance, it is easy to understand that in a production process, the quantity and quality of a good produced is dependent on the quantities of various inputs produced. In addition, there may be external inputs that, although not chosen by this producer but by someone else, may have a negative or positive impact on the production of the subject product. As in the example of negative externality from farmland cultivation, the total fish production in the downstream lake is not only dependent on fishing inputs employed by the lake users. The fish production is also affected, in this case negatively, by the levels of sediments and chemical residues deposited through the water runoff from the upstream farm. Similarly, the consumption utility or value derived by each cell phone user in the network depends on the number of phones purchased by other users.

Another important aspect of the externality that can be ascertained from the above definition is that the spillover or side effect of one person's action on the other is *not deliberate* or is *unintentional.* An economic agent who is responsible for a given externality has no intention of causing harm or doing good to the other party. The original action is undertaken for one's own personal gain (e.g., appropriating profit or income from crop production or enjoying certain utility from using the cell phone). The negative or positive spillover effect of such action is strictly unintentional.

Suppose that you are sitting in a room or a restaurant next to a person who is smoking. You may feel uncomfortable with the smell or aesthetic of the smoke. You may also develop a health problem if you are exposed to that smoke over a long time. As long as smoking is allowed in the room, the smoker is engaged in a legally acceptable activity. However, the secondhand-smoke effect inflicted on you is a negative externality. There is physical connection between the smoker and you. On the other hand, let us say every time when you sit near that smoker, he turns around and deliberately blows a puff of smoke right on your face. The effect of this action may be the same as in the first case. This deliberate action, however, would not constitute an externality because it is unintentional. Instead, we would call this behavior a nuisance or even a crime if that person does it every time he sees you.

The above definition covers a specific type of externality. These are external effects that fail to transmit through the market system in the form of prices. When a producer does not account for the external costs she inflicts on a third party, the price that she is willing to accept does not reflect the social costs of production but only her private costs of production. The market prices of such products will be cheaper than otherwise. This is because *no* private market contract exists between this producer and those affected by external effects that would have required the producer to pay for such external costs. This applies to our

cell phone market example also. When a new buyer buys a cell phone, he enhances the value of phones in use for all the existing network users. The seller does not have any ability to go back to the existing customers and to ask for a cut in the additional value accrued to each of them. The seller normally may increase the price on future cell phones sales, knowing that her product is more valuable to customers than before.

In both the examples above, the externalities are caused by real variables (e.g., sediments or phone), the economic (monetary) effects of which fail to transmit through the market process. In the case of negative externality, the full economic effects (costs) of production do not reach the intended buyers (i.e., consumers of farm products) in the market. Instead, a portion of the total costs (i.e., off-farm pollution effect) is felt by a third party (i.e., lake users). Such externalities are called technological or *nonpecuniary externalities*. In economics, the inability of the market to capture the full social costs and social benefits, therefore sending the right price signal to producers and consumers of a given commodity, is referred to as *market failure.*

The following modified definition of externality (Lesser, Dodds, & Zerbe, 1997) captures the essence of the above discussion. A technological externality is *an unintended side effect of one or more parties' actions on the utility or production possibilities of one or more other parties, and there is no contract between the parties or price system governing the impact.*

The above definition excludes economic side effects that manifest through market prices. For instance, a city municipal government builds a landfill in one of its neighborhoods. The noise and the ugly look of trucks carrying the city solid wastes daily around the landfill disturb the residents. The foul smell from the landfill may become a constant reminder of the landfill's presence. Those homeowners who can afford to buy houses in better neighborhoods elsewhere may want to sell their homes. As more and more homes start appearing in the market and new buyers may not prefer to purchase a house near the landfill, the home prices in the area will decline. This loss of home value is certainly an unintended economic side effect caused by the city's decision to build the landfill. However, this effect is carried through the housing market price. Such external effects are referred to as *pecuniary externalities.* While some people may argue that the loss in the property value is unethical or an injustice, this loss does not represent a *market failure* problem. In fact, in this case, the housing market seems to have done its job. That is, the market accurately has reflected the lower preference that the new buyers have for homes near the landfill, or that the buyers are willing to pay lower prices.

### *Effects of Externality*

As hinted earlier, the presence of externalities causes a private market to fail. When all the conditions of a perfectly competitive private market are fulfilled, the market is assumed to deliver most efficient results. Profit-maximizing producers allocate their scarce resources in a most efficient fashion. That is, the total quantity of their output meets the condition of marginality: The marginal benefit (i.e., the market price in the case of perfect competition) is exactly equal to the marginal costs of production. Similarly, the amount that consumers are willing to pay (the marginal willingness to pay) balances with the market price offered to them at the margin. This is the market equilibrium level. At such point, the market equilibrates in the sense that the producers' marginal costs of production are identical to consumers' marginal willingness to pay (marginal benefit), which is further equal to the market price. At this point, the combined net surplus of producers and consumers is maximized. Both producers and consumers are assumed to have made their resource allocation decisions in a most efficient fashion. This point of market equilibrium is therefore called *socially efficient* equilibrium.

The above market outcome is based on certain assumptions. The marginal costs of producers (i.e., supply curve) must represent the total or social costs of production. The marginal willingness to pay of consumers (i.e., market demand curve) must represent the total benefit that they derive. That is, the market supply curve should have taken into account both private costs (representing the opportunity costs of all production inputs and the production technology used) and all the potential external costs. On the other hand, the market demand curve should reflect consumers' private benefits (their own preference and taste) and the external benefits. As we discussed earlier, if there exist external costs and benefits, the private market is not going to capture these additional effects. The resulting market outcome (equilibrium quantity and price) does not reflect the social (full) costs and benefits either. It only captures the private costs and benefits and therefore can only be viewed as a *private market efficient* equilibrium.

Going back to the problem of agricultural runoff, let us assume that the private farmer faces a downward-sloping market demand, $D$ (Figure 22.1[a]). The demand curve represents the aggregate marginal willingness to pay of all consumers in the market, *MWTP.* The farmer has a marginal private cost function (supply curve), *MPC.* Assume that every unit of agricultural output imposes certain marginal external costs, *MEC*, on the lake users. Thus, the agricultural production comes at marginal social costs, *MSC*, to society. Note that

$$MSC = MPC + MEC.$$

In the private market, where she has no obligation to consider the external costs, the farmer produces $Q_m$ units of output at a unit price of $P_m$ (i.e., at a point where $MPC = MWTP$). This is efficient only from the view of the private producer, and therefore we call this level private market efficient equilibrium ($Q_m$, $P_m$).

The above equilibrium, however, is not socially efficient because the underlying cost function does not include the external costs. The socially efficient equilibrium ($Q_*$, $P_*$) should occur at a point where $MSC = MWTP$. It is interesting to note two points here. The socially efficient equilibrium quantity is lower than the private market efficient equilibrium quantity. Also, the socially efficient equilibrium price is higher than the private market efficient equilibrium price. The private market oversupplies the good. That is, $Q_m > Q_*$ and $P_m < P_*$. A production in excess of socially efficient production means larger pollution loading downstream. Because the external costs are left out of the producers' cost equation, consumers will enjoy a lower market price.

Now, consider the cell phone example that involves external benefit. Assume that the cell phone company faces a supply or marginal private cost function, *MPC* (Figure 22.1[b]). We assume that there is no external cost resulting from the phone production. All cell phone users face a combined private marginal willingness-to-pay curve, $MWTP_p$. That is, this relationship includes only the private benefit portion of all network users. We know that the each user (or phone) adds certain external value to all customers, $MWTP_e$. Thus, the (total) social marginal willingness to pay of each additional cell phone is $MWTP_s$:

$$MWTP_s = MWTP_p + MWTP_e.$$

We can see that in this case, the private market produces only $Q_m$ at a private market equilibrium price of $P_m$. However, because of the presence of external benefits, the socially efficient market equilibrium occurs at $Q_*$ and $P_*$. Note that, unlike in the case of external costs, when there are external benefits, the market underproduces goods and services (i.e., $Q_m < Q_*$). Because the consumers' true preference or value is not fully revealed in the marketplace, the private market equilibrium price is lower than the socially efficient equilibrium price (i.e., $P_m < P_*$). One might argue here that consumers are better off with lower market price. That is true; however, the network as a whole will end up getting a smaller number of cell phones than what is socially optimal if the producer is not in a position to capture the full value of each marginal phone supplied in the market. This is particularly a problem in a special class of market goods, called *public goods,* which we will discuss later.

### *Causes of Externality*

The question that naturally arises is why externalities exist in the market. In the previous examples, we see that the producer has no incentive to compensate the victims of downstream externalities, nor do cell phone consumers have an incentive to pay for the external benefits they enjoy from each additional network user. In other words, the economic decision makers in each case have no incentive to *internalize* the external costs or benefits. The two most fundamental reasons of externalities are (a) nonexclusivity and (b) high transaction costs.

First, externalities are often associated with nonexclusion or nonexclusivity characteristic of a resource or an economic asset. Consider again the example of our upstream–downstream agricultural pollution problem discussed earlier. There are no legal, social, or technological barriers that prevent the farmer from loading sediment and chemicals into the downstream lake. That is, the farmer has

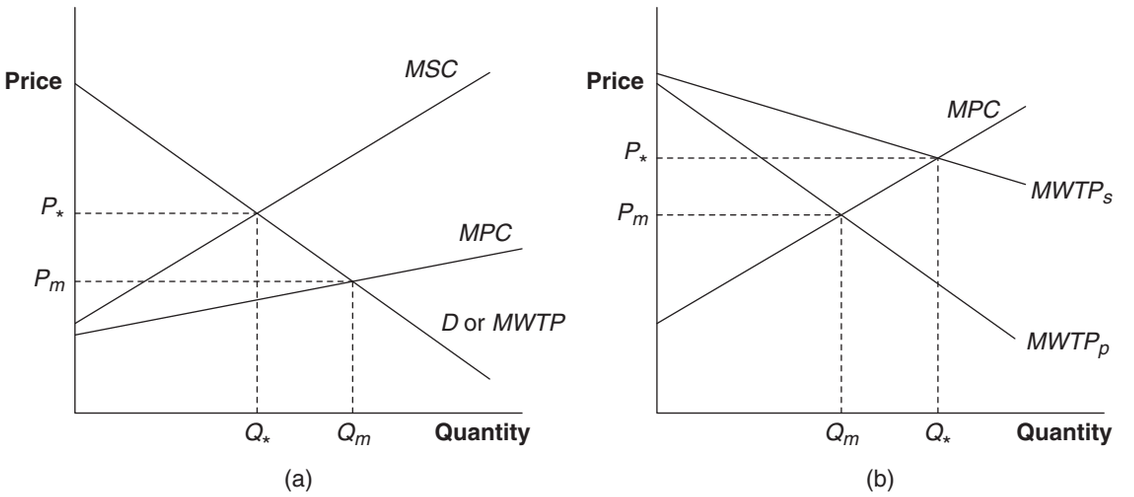


**Figure 22.1** Private and Socially Efficient Market Equilibria: (a) The Case of Negative Externalities; (b) The Case of Positive Externalities

uncontrolled access to the lake. Under the existing legal or social norm, the lake users cannot exclude (nonexclusivity feature) the farmer from dumping pollutants into the lake. The ownership of the lake is not clearly defined. Similarly, in the case of the cell phone example, as each new buyer enters the market, the existing customers benefit from the additional network sociability or social interaction. There is nothing in the marketplace that prevents the existing customers from enjoying that extra benefit unless the phone company decides to switch off their connectivity. However, such practice might be too costly for the phone company. Again, this is a case of nonexclusivity.

This brings us to the second condition of efficient markets, mentioned in the beginning of the chapter: *well-defined property rights.* The degree to which the users of a resource or property can exclude others depends on the nature of *property rights.* The right to own and use a property, including the right to exclude others from using it, influences whether and how much the external costs and benefits are internalized and transmitted through the market price. In the next subsection, we will explore the role and significance of property rights in more detail. At this point, it is enough to note that one of the reasons the lake users are forced to bear the burden of the negative externality is that they lack a well-defined property right to the lake and, therefore, lack the right to exclude upstream farmers from dumping pollutants to the lake.

The second reason externalities exist is the high *transaction* costs. According to Arrow (1969), transaction costs are *the costs of running the economic system.* Even when the right of exclusion exists, the property owners may not be able to exercise that right due to high costs of transaction. The transaction costs include the costs of gathering necessary information, costs of reaching an agreement with the other decision makers in the market with regard to one's right, and the costs of enforcing such agreement. For instance, in the case of lake users, the costs of proving that the pollution loading originates from the upstream farmer may be high. This is especially true if the pollution problem is not from just one particular farmland (a point source pollution) and, instead, from a large number of farm lands (a non–point source pollution). There are additional costs of reaching a private agreement with the upstream farmers and making sure that they bear the full costs of the pollution or take action to abate pollution. Basically, the transaction costs are the costs of exclusion. The lack of, or ill-defined, legal right to exclude upstream farmers means that the downstream lake users have high transaction costs of exclusion. These costs may be too high for downstream users to try enforcing their right against upstream farmers.

The amount of transaction costs depends on the technical and physical nature of the externality and the number of parties involved in a given market transaction. In many cases involving negative or positive externalities, there may be technical or physical barriers in reaching an agreement. Certain pollution may be hard to detect with the available technology for measurement. The external costs of such pollution may not show up in the short term but only in a distant future. A precise estimation of such costs is often difficult. External benefits may pose similar problem. Putting a value on the network-wide aggregate incremental benefit is not easy in the case of the cell phone example. Obviously, the larger the number of parties involved on either side of a market transaction, the higher the transaction costs of gathering information, reaching the agreement, and enforcing the agreement.

Whether a producer who is responsible for certain external costs is willing to bear or *internalize* the externality depends on whether she is able to transfer that extra cost on to her consumers. For her to do so, the market price that the consumers are willing to pay must be higher than the private costs of producing the good and the external costs. If the market price is not high enough to cover both the private costs and external costs, the producer has no option but to ignore the latter, which she can easily do because she enjoys nonexclusive access to the downstream environment. Thus, the externality comes into existence when the costs of internalization exceed the gains from internalization.

## Property Rights

According to Demsetz (1967), property rights are an instrument of society that accords rights to an individual to own, use, dispose, sell, and responsibly manage certain property or economic assets. These rights may be expressed through laws, social customs, and mores. The rights give its owner an uncontested privilege to appropriate benefits from the designated property or asset. The rights also protect the owner from other individuals of the society from encroaching on this asset. The bundle of rights normally comes with certain responsibility so that the owner may not put the property to legally and morally unacceptable uses.

A clearly defined property right lends its owner the ability to appropriate full benefit from its use or management. For a private market to operate efficiently, this is a key requirement. When a producer makes his decision to invest in factors of production (land, labor, and capital), he forms an expectation that he gets the full reward for his effort. One of the necessary conditions that make this possible is that he possesses full rights to all the factors and the final product until he decides to dispose his right to that product in the marketplace for a due compensation (i.e., a reasonable market price). Such a guarantee of right gives the owner an incentive to invest in the asset, manage it in a most efficient fashion, and dispose of it only when he thinks it is the best time. Therefore, the conventional economic wisdom holds that privately owned resources are put to most efficient uses. If all privately held scarce resources are used efficiently, the society as a whole benefits from that process as well.

In order for property rights to serve their function well in the market, the following requirements must be met:

Property rights must be *well defined.* There should be no ambiguity as to who owns a given economic asset. The clear identification of ownership gives its owner a right to make decisions relating to the asset's growth, maintenance, use, and disposal. The ownership can be *private, communal,* or *state.* A private owner may be an individual or a firm. The car you legally own is an example of private property rights. The farmland owned by the farmer is another example of private property rights. A patent right owned by a private firm to a certain technological innovation or product is also a private property right. A communal ownership is normally a group of owners, each member of which will have certain use and access rights and responsibilities. Communal ownerships are common among traditional societies in the case of forests, grazing land, water, rivers, and other natural resources. A state ownership puts the resource or asset in the control of a government agency. The state may own and use the resource exclusively or may share a portion of the rights with its constituents or citizens. For instance, a national park in the United States is owned by the U.S. government and managed by the National Park Service. The parks' certain uses (recreation, research, or mining in the park) may be open to the general public, subject to certain rules and regulations.

Property rights must *exclusive.* The owner of the property rights must be able to exclude others from using it. This exclusivity lends the owner an ability to enjoy the full benefit from it and be able to recover the entire costs of operating it. As discussed in the earlier sections, exclusivity also prevents an external individual from causing harm to the owner (i.e., externality).

Property rights must be *enforceable.* There must be clear legal or social guidelines as to how the society deals with a violation of one's property rights. If somebody encroaches upon the farmland, the farmer must be able to bring the violator to justice, meaning that the latter is evicted from the property and made to pay the cost of evictions and any economic damage associated with the encroachment.

Property rights must be *transferable.* The owner must be able to dispose the property whenever he wants and at a price that he can reasonably get from the market, without much restrictions. If the property is divisible and the owner can dispose any portion of it as needed, such property right represents full transferability. The owner can more effectively manage his or her resource.

### *The Case of Weak Property Rights: Open-Access Resources and Public Goods*

We will find many instances in the market system where economic resources do not operate under strict private property ownerships. In fact, we already have introduced several examples of these resources. When the property rights are ill defined, weak, or absent, multiple users will claim ownership or access rights to such resources. For instance, both upstream farmer (though unintentionally) and downstream lake users can access the lake and have competing uses for the same. Cell phone network users can enjoy the additional benefits generated by each additional phone buyer. The network provides a certain value component (increased sociability) that is open to all users to enjoy without excludability. Thus, when the property rights cannot be assigned or defined, the problem of nonexcludability arises. As we said earlier, nonexclusivity is one of the determinants of externality and market failure.

Two types of natural and economic resources and services suffer the problem of nonexcludability: *open-access resources* and *public goods.* These two resource types differ in the way their users interact with each other. We will discuss each of these resource types individually and the unique problems they pose for their owners and society in general.

An open-access resource is that resource to which users have uncontrolled access, leading to competition among users. That is, the consumption or extraction of the resource by one user reduces the quantity available for others. Thus, the resource is *divisible* and *subtractable.* In other words, you will find *rivalry* among its users. Thus, open-access resources are nonexclusive, rivalrous in nature. Fishery resources in open international waters are a good example of open-access resources. You may not find many strictly open-access resources today. A large number of resources are placed under government or community control. However, for all practical purposes, their users may have open and unlimited access to them. Such resources do suffer to some degree the problem of nonexclusivity and competition among users. Grazing lands, rivers, government forests, fishing grounds, national highways, and public parks are some examples.

Users of an open-access resource basically lack incentives to exercise caution or restraint in its use. Because other users cannot be excluded, a given user trying to conserve or enhance the resource may not realize the full benefit of his action. Therefore, the common motive you observe among the users of open-access resources are "use the resource while it lasts," "grab the resource as quickly as you can," and "grab as much of the resource as you can." Such competitive behavior will lead to overexploitation of the resource. This overexploitation problem is what Hardin's (1968) famous expression, the "tragedy of the commons," refers to. Each user will try to maximize his or her own self-interest (i.e., to the point where the private gains from the additional unit are equal to the private costs of the additional unit). In this process, each user may negatively affect the productivity of the resource or costs of extraction of the resource for all other users. The impact of each user's action on oneself may be insignificant, but the

effect of one user's action on all users combined could be quite substantial. In Hardin's words, the "freedom in a commons brings ruin to all."

To illustrate the above problem, let us consider a common-pool ground water aquifer. Farmers who own land on the ground directly above the aquifer can easily access the aquifer and extract water. Because it is accessible by only those who own land above it, the aquifer is normally considered as a common-pool resource (i.e., a resource owned by a definite community of people) and is not strictly an open-access resource. However, those land owners will have uncontrolled freedom of access to it and therefore could construct as many wells as they need.

Suppose that each well can serve two acres of land and generate a total gross return of $1,000, net of all other but the pumping cost. The cost of pumping is an increasing function of the pumping depth, which in turn is an increasing function of the number of wells on the aquifer. That is, as the number of wells increases and more water is withdrawn, the water table goes down and the pumping lift and the cost increase. See Table 22.1 for the costs of pumping per well. Looking at the net revenue column, you will see that the net revenue per well gradually declines and reaches zero when the number of wells is 60. The aggregate net revenue from all operating wells combined, though, is maximized at 30. This level can be considered the socially efficient level of extraction ($Q_*$). However, because the landowners have uncontrolled access to the aquifer, their tendency is to dig wells as long as each additional well brings a positive net profit (i.e., 60 in this example). This level can be considered the private market solution ($Q_m$). Note at the market equilibrium level, the aggregate profit earned from the resource is zero, an indication of most inefficient economic overexploitation.

Now we will turn to the case of public goods. Like open-access resources, *public goods* also suffer a nonexcludability problem. The difference is that there is no rivalry among the users of public goods. The same quantity of the good made available to one person is automatically available to all users. That is, users jointly consume or enjoy the good. This refers to the *indivisibility* quality of the good. In the words of economist Paul Samuelson (1954), "Each individual's consumption of such a good leads to no subtractions from any other individual's consumption of that good" (p. 387). Common examples of public goods are clean air, open landscape, biodiversity, public education, national security, a lighthouse along the beach that the mariners use, flood protection services, and ecological services. The word *public* here means that it is available to all users (the public) at the same time and at the same quantity.

There can be privately provided public goods. For instance, radio signals or network television services provided by private companies are also public goods, although they are paid for by private companies. Who (private or public entity) pays for a good does not make it a "public good," but that it is made available for the general public in lump sum does.

An important feature of public goods is that they possess a large degree of external benefits. A few mariners may decide to invest in building a lighthouse so that they can safely navigate around the coast. However, this lighthouse can be easily used by those mariners who may have not paid for it. Thus, the efforts of those invested in the lighthouse are yielding external benefits to noninvestors also. The nonexclusivity problem makes it difficult for the original investors to recover compensation from the noninvestors. The costs of excluding the latter may be exorbitant.

The nonexcludability (lack of property rights) and the presence of external benefits pose a unique problem in the case of public goods: the problem of *free riding.* Free riding refers to enjoying the benefits of others' efforts without paying for them. Once a public good or service is provided, users have no incentive to contribute toward the cost of the service. In a private market, you may not be able to find a sufficient number of buyers who can pay enough to make such services or goods available at socially efficient levels—that is, the problem of undersupply as discussed in the case of positive externality (Figure 22.1[b]), $Q_m < Q_*$. Again, this is another case of market failure, caused by the presence of weak property rights and positive externalities.

**Table 22.1** Income and Costs From an Open-Access Groundwater Aquifer

| *Number of Wells* | *Revenue per Well ($)* | *Pumping Costs per Well ($)* | *Net Revenue per Well ($)* | *Total Net Revenue From the Aquifer ($)* |
|---|---|---|---|---|
| 10 | 1,000 | 100 | 900 | 9,000 |
| 20 | 1,000 | 200 | 800 | 16,000 |
| 30 | 1,000 | 400 | 600 | 18,000 |
| 40 | 1,000 | 600 | 400 | 16,000 |
| 50 | 1,000 | 800 | 200 | 10,000 |
| 60 | 1,000 | 1,000 | 0 | 0 |

## Policy Implications

The presence of externality and ill-defined property rights lead to market inefficiency (i.e., either undersupply or oversupply of certain goods and services). There are several policy options for correcting this problem. These policy options fall under two broad categories: property right solutions and government interventions.

### Property Rights Solutions

We had argued that one of the reasons for the presence of externalities was the absence of well-defined property rights and the attendant nonexclusivity. So to resolve the externality problem, economists suggest that we directly target the weak property rights. This approach is a private market solution or property right method. Nobel laureate economist Ronald Coase (1960) articulated this approach, which is now popularly known as the Coase Theorem. The essence of the theorem is simple: If a clear property right over an environmental asset involving externality is specified, a negotiation between the two private parties (originator and the victims of the externality) will emerge and eliminate the externality. Irrespective of who originally has the property right over the environmental asset, a private market for trading the externality will evolve.

With the lake example, the private right over the lake may be given to either the upstream farmer or downstream lake users. If the farmer has the right, it makes sense for the lake users to enter an agreement with the farmer to have her cut back the pollution. Lake users will have to compensate the farmer toward the pollution control costs. This of course is based on the assumption that such monetary compensation is lower than the gains from the avoided damage to the lake via pollution control. Similarly, if the lake users have the property right, the farmer will have to pay suitable compensation to the lake users to accept a certain amount of pollution and associated damage. Thus, in either case, a more efficient private market outcome will result, with reduced pollution and lower external costs. What is happening here is that the clearly defined property right forces the parties involved to *internalize* the externality.

The above market solution will work only under certain conditions (Field & Field, 2009). The transaction costs of establishing and enforcing the market agreement between the parties must be reasonably low. A higher transaction cost might discourage either party from entering into an agreement. If the number of parties involved is large or it is too hard to prove the pollution damage impacts, the transaction cost is normally high. With a large number of parties involved, there is also a possibility of a free-rider problem. Furthermore, for technical, social, and political reasons, in some cases it may be hard to establish clear property rights.

Political scientist Elinor Ostrom (1990), who received a Nobel Prize in economics in 2009, argues that communal property rights can also eliminate overexploitation and reciprocal externality in the case of open-access resources. The *Ejidos* system of community forest management system in Mexico is a good example of a communal property right system. Almost 8% of the country's forest lands are owned and managed by community organizations. Rules of access and profit sharing are formed and enforced by the communal body and recognized by the government. Community-managed *sacred groves* found in other countries serve a similar role but on a more informal basis. A similar group or community approach is possible in the case of public goods. For instance, home owners form an association or club to manage common facilities in their residential area (e.g., playgrounds, recreational facilities, lake). Homeowners will then be required to pay for these common services with a monthly fee or on a pay-per-use basis. Such public goods and services now become *club goods*.

Finally, government may take open-access resources into its control and place full or partial restrictions on access and use rights. The goal here is to eliminate competition among users and possible externality imposed on each while consuming the resource. In the case of public goods, in the overall interest of the society, the government may bear the full cost of supply and operation instead of relying on private markets. National defense, public radio and television services, and public education are some examples of this approach.

#### *Government Interventions*

Privatizing open-access resources and government provision of public goods may not work effectively in all situations. We saw a number of reasons why it may be hard to establish property rights over resources and thereby internalize externalities. In those cases, the government can directly intervene into private markets using a variety of policy instruments, to influence the production and consumption behavior of economic agents. There are two types of government interventions: command-and-control policies and market-based policies.

The command-and-control policies will impose legal restrictions on producers' and consumers' behavior. Formal legislation may be enacted by local, state, and federal governments specifying whether and how much pollution is allowed. Private factories spitting smoke into the atmosphere and discharging waste into water bodies may be asked to install certain pollution abatement devices at their site. These are called *technology standards.* Or, the government may set a maximum limit or performance standard on the quantity of emissions leaving their plants during each time period, which may be achieved through any means of the polluters' choice. These are called *emission standards.* In the case of open-access resources such

as fisheries, ground water aquifers, and forests, the government may put legal restrictions on quantity extracted, timing of extraction, and nature of extractions. Automobile owners may be required to install certain emission control devices on their cars. These command-and-control policies require proper enforcement mechanisms, including provisions for a penalty on those who fail to comply with the law. This approach is justified on the *polluter pay principle* and is viewed as a fair approach by many, including environmentalists. While this method is easy to administer from the government agency point of view, it may work out to be a more expensive way of internalizing externalities.

Market-based approaches to internalizing externalities may overcome some of the limitations of command-and-control approaches. Again, there are several options: taxes, subsidies, and tradable permits. These approaches give economic agents higher decision flexibility, make them weigh the costs of creating an externality (e.g., an emission tax) and the costs of not doing so (e.g., emission abatement), and help them choose the most cost-effective means of internalizing the externality. Under taxes, producers responsible for a certain externality will be required to either pay a tax on every unit of pollution (emission tax) or production output (output tax) causing that externality or invest money in pollution control. Under this policy, the producer will have internalized the externality either by lowering the emission and therefore bearing the costs of abatement or by paying taxes on uncontrolled emission, which may be used to compensate the victims. A subsidy is another approach whereby government will encourage socially desirable behavior (pollution control or reduced resource harvesting) among private economic agents.

Tradable emission permits and harvest quotas are more recent and innovative market approaches that create private rights over emission or resource consumption. The popular cap-and-trade system for greenhouse emissions or sulfur dioxide emission is an example of this policy. Under this policy, each polluting firm will have a certain number of permits, each permit being equal to the right to emit a unit of pollutant. The firm can use all the allotted permits, save some and sell the unused permits in the market, or purchase some permits from other permit holders. Each firm will compare its own marginal costs of abatement with the market price of the permit. Those firms whose marginal costs of abatement are higher than the permit price can save money by purchasing permits in the market and, therefore, by not having to lower the emission as much. On the other hand, firms whose marginal costs of abatement are lower than the permit price can make a profit by lowering their emissions and by selling the unused permits at a price that is higher than their marginal abatement costs. After trading, all the firms in the market (buyers and sellers of the permits) will have adjusted their respective emission levels such that their marginal abatement costs are made equal to the permit price and, therefore, are balanced across all sources. This phenomenon is called the *principle of equimarginal costs.* In this process, the permit trading system will have moved the emission control responsibility in the market from high-cost (less cost-efficient) firms to low-cost (more cost-efficient) firms. As a result, the overall costs of emission control (i.e., costs of internalizing the externality) will be minimal.

Individual transferable quota programs for fisheries and water trading between urban and agricultural water users are other examples that fall under this category. A long-term advantage of this policy approach is that the economic agents will have an incentive for technological innovation. The permit or right system induces them to invest in innovative and cost-effective means of pollution control or harvest methods. Those who innovate find that they do not require all the allowed permits and rights and can sell the unused ones in the market for a price.

## Conclusion

In this chapter, we learned that the presence of externalities and weak property rights renders private markets inefficient. When economic agents cause negative externalities, the market will produce above what is socially optimal or efficient, leading to wastage of production resources, excessive output, and undesirable external impacts. Negative externalities associated with production or consumption will go unpriced in the market. On the other hand, in the case of positive externalities, the market will underproduce because consumers may not pay for the external benefits created as a result of someone else's action.

Externalities and property rights are related concepts. The primary reasons for externalities are nonexclusion and high transaction costs. Nonexclusion arises from weak or absent private property rights over resources or market services. Open-access resources and public goods are two common examples where weak property rights prevail and externalities emerge.

The presence of externalities and weak property rights calls for government intervention. One approach is to fix the weak property rights, which of course may not be possible in all cases. Properly defined property rights will force private agents to internalize externalities, moderate their consumption and production levels, remove undesirable conflict among users (open-access resources), and eliminate free-riding incentives (public goods). Direct government intervention, with different degrees, is another popular means to correcting externalities. No one size fits all.

## References and Further Readings

Arrow, K. J. (1969). The organization of economic activity: Issues pertinent to the choice of market versus nonmarket allocation. In U.S. Congress, Joint Economic Committee (Ed.), *The analysis and evaluation of public expenditures: The PBB system* (Vol. 1). Washington, DC: Government Printing Office.

Baumol, W. J., & Oates, W. E. (1988). *The theory of environmental policy.* Cambridge, UK: Cambridge University Press.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics, 3,* 1–44.

Demsetz, H. (1967). Towards a theory of property rights. *American Economic Review, 57,* 347–359.

Field, B. C., & Field, M. K. (2009). *Environmental economics: An introduction.* Boston: McGraw-Hill Irwin.

Hardin, G. (1968). The tragedy of the commons. *Science, 168,* 1243–1248.

Lesser, J. A., Dodds, D. E., & Zerbe, R. O., Jr. (1997). *Environmental economics and policy.* Reading, MA: Addison-Wesley.

McMillan, J. (2002). *Reinventing the bazaar: A natural history of markets.* New York: W. W. Norton.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action.* Cambridge, UK: Cambridge University Press.

Samuelson, P. A. (1954). The pure theory of public expenditure. *Review of Economics and Statistics, 36,* 387–389.

# 23

# Public Choice

Peter T. Calcagno
*College of Charleston*

Public choice economics is the intersection of economics and politics. It uses the tools of economics to examine collective decisions. Public choice economics reflects three main elements: (1) methodological individualism, in which decision making occurs only with individuals; (2) rational choice, in which individuals make decisions by weighing the costs and benefits and choosing the action with the greatest net benefit; and (3) political exchange, in which political markets operate like private markets, with individuals making exchanges that are mutually beneficial (Buchanan, 2003). It is this last element that makes public choice a distinct field of economics.

Using the basic tools and assumptions of economics on collective decisions provides many insights. Economists assume that individuals are rationally self-interested in private decision making. However, this assumption is not always applied to the realm of collective decision making. In fact, for the first half of the twentieth century, it was argued that collective choices are made based on what is best for society. This view is referred to as *public interest theory*. Early public choice scholars called this dichotomy of behavior into question, arguing that individuals are rationally self-interested regardless of which sector they operate in. The public interest theory does not allow for someone operating in the public sector to make decisions based on what would benefit solely him or her. However, the public choice theory, properly understood, does acknowledge that when an actor in the public sector acts in his or her own self-interest, it may be consistent with the public's interest. In this regard, the public choice view is a more encompassing theory.

Assuming that individuals are rationally self-interested and applying the laws of supply and demand provides a unique understanding of political decision making. (For the purposes of this chapter, the terms political and collective decision making are used interchangeably.) Voters can be thought of as demanders or consumers of public policy; as such, they are concerned with having policies enacted that will benefit them. Politicians act as suppliers of public policy. Just as businesses in the private sector compete for consumers, politicians compete for voters. If businesses want to maximize profits, then politicians want to maximize votes. Unlike in private markets, public sector decision making has a third party involved in the decision making process: bureaucrats or civil servants. These are the individuals who run government and carry out the public sector choices and policies on a day-to-day basis. Bureaucrats are not elected, which means they do not directly serve a group of constituents. The public choice view of bureaucrats is that they maximize power, prestige, and perks associated with operating that bureau. To accomplish this goal, bureaus maximize their budgets.

Public choice theory provides a way to evaluate collective decision making that not only allows for positive analysis but also addresses some of the normative issues associated with policy making. The rest of the chapter provides an overview of the origins of public choice theory and the contributions it has made to the discipline of economics.

## Origins of Public Choice

Public choice economics emerged from public finance, which is the analysis of government revenues and expenditures. According to James Buchanan (2003), it was

evident by the end of World War II and the early 1950s that economists did not have a good understanding of public sector processes. Economists began to realize that the naive public interest view did not reconcile with the reality of politics. Buchanan made his first endeavor into this issue in 1949. Around that same time, two other scholars began to examine a similar question. Duncan Black (1948) began examining the decision-making process of majority voting in committee settings. Kenneth Arrow (1951) was examining whether choices of the individual could be aggregated to create an overall social ordering of preferences. Arrow concluded in what is now known as his *impossibility theorem* that it is not possible to aggregate preferences to develop a social welfare function. The only way the preferences of individuals could be consistent with those of society is if a dictatorship existed and the preferences of society were those of the dictator's. Arrow's theoretical contributions served to inspire public choice theorists, but his contributions formally developed into what is now known as *social choice theory.* Social choice theory is focused on the concern of social welfare and the public's interest. It examines how collective decisions can be made to maximize the well-being of society. Further analysis of social choice theory is beyond the scope of this chapter (see Arrow, Sen, & Suzumura, 2002, for more on social choice).

Public choice theory was developed primarily on the work of Buchanan, who was awarded the Nobel Prize in Economics in 1986. Buchanan was inspired by Swedish economist Knut Wicksell, who can be considered a precursor to public choice theory. Wicksell's focus on the role of institutions led Buchanan and Gordon Tullock (1962) to write *The Calculus of Consent: Logical Foundations of Constitutional Democracy.* This seminal work is considered the origin of the public choice revolution (Gwartney & Wagner, 1988).

The public choice revolution that began in 1962 with *The Calculus of Consent* called into question not only how political decisions are made but also the role of government in the private sector. Prior to the development of public choice theory, many economists argued that there are instances when private markets will fail to allocate resources efficiently: externalities, public goods, and monopolies. The conventional view was that government intervention could improve the allocative efficiency through taxes, government production, or regulation of the market. Public sector officials were thought to make decisions based on the public's interest and the improvement of the overall welfare of society. However, this view is at odds with the economic and political reality of public sector outcomes. While most economists do not favor policies such as price controls, tariffs, and regulatory barriers that impede competition, these types of policies are common public sector outcomes. More specifically, public choice economists started asking why, if government intervention is the means to correct market failure, does government enact policies that often increase economic inefficiency. Public choice's concept of all individuals acting in rational self-interest provided an answer to this question. It reminded us that government is not an entity that is seeking social welfare but is merely a set of rules under which decisions are made. While self-interest leads individuals to socially enhancing outcomes, via Adam Smith's invisible hand, these same motivations under government institutions can actually create outcomes that are not beneficial to society. The conclusion that emerges from this field of economics is that if one is to acknowledge that market failure exists, one must recognize that government failure exists.

## Voting in a Direct Democracy

The issue of voting and majority rule was at the forefront of the development of public choice theory, and it is still the subject of much of the literature written today. The numerous aspects of voting addressed by public choice scholars are too vast for this chapter; therefore, the focus is on the seminal work in this area.

### Majority Voting Cycles

Beginning with the work of Black (1948) and Arrow (1951), the idea emerged that if preferences are not single peaked when individuals vote for two candidates, referendums, or policies, majority voting can result in a cyclical voting pattern. This concept can be illustrated with Table 23.1. Three voters, Gordon, James, and Anthony, are represented across the top of the table, with each voter's ranking of preferences represented by the order in the column. Thus, column 1 shows that Gordon ranks his preferences as A, B, and C. Similarly, columns 2 and 3 show the preferences of the other two voters. Based on the preferences of these three individuals, there is no decisive winner when these options are voted on in pairs. Option A paired with B will result in A winning. Option B will defeat Option C, and similarly, Option C will defeat Option A. The outcome will be determined by the number of elections that will occur and in what combination the options are paired. There is not a single option that wins the majority of votes. Two further concepts regarding voting, developed in light of voting cycles, are the median voter model and agenda manipulation.

#### *Median Voter Model*

The median voter model arises out of the need to assume single-peaked preferences. Single-peaked preferences exist when voters prefer one option above all other options. If single-peaked preferences exist, then cycling is no longer a concern. In addition, the option must be a single-dimensional issue, for example, measuring government expenditures from left (less) to right (more) on an x-axis. With these assumptions, a majority rule will lead to the median voter's position always winning. The median

**Table 23.1** Preference Analysis for Three Voters

| *Gordon's Preference Ordering* | *James's Preference Ordering* | *Anthony's Preference Ordering* |
|---|---|---|
| A | B | C |
| B | C | A |
| C | A | B |

voter model has several applications for evaluating voting decisions: committees, referenda, and representative democracies.

Figure 23.1 demonstrates the median voter model. Imagine that Figure 23.1 represents spending on education to be voted on by a committee. The price axis (P) represents the amount of education spending. Committee members Gordon, James, and Anthony all pay the same tax price (T) $500. The difference is the quantity they would like to receive for that tax price. The horizontal or quantity axis (Q) represents the amount of education each voter prefers. Gordon prefers quantity $Q_1$, which is less than James, who is the median voter and desires $Q_M$, and Anthony prefers the largest quantity, $Q_3$. Given the voters' single-peaked preferences, the single dimension of the issue, and either a greater or lesser quantity of education at T, then in a majority vote, $Q_M$, the quantity preferred by James (the median voter), will always win. In a committee environment, Gordon would suggest $Q_1$, his most preferred outcome; then James, the median voter, would counter with $Q_M$, his most preferred outcome. Gordon would vote for $Q_1$, while James and Anthony would vote for $Q_M$, and $Q_M$ would win by a vote of 2-to-1. If Anthony proposes $Q_3$, then James would propose $Q_M$, and this quantity would again win by a vote of 2-to-1, Gordon and James having voted for $Q_M$ and Anthony having voted for $Q_3$. Any pairing with $Q_M$ will result in $Q_M$ winning because $Q_M$ would be preferred by Gordon and Anthony over $Q_3$ and $Q_1$, respectively (Holcombe, 2005).

This proposition will also hold for a referendum, where the government would propose a quantity of education (still referring to Figure 23.1). If the referendum were to fail, another one would be proposed until the amount $Q_M$, which could win by a majority, was on the ballot.

The median voter model, when applied to selecting a candidate, is a slightly different analysis. Voters decide on a candidate who will represent their preferred position on a multitude of issues—that is, a party platform. Figure 23.2 demonstrates the median voter model in the context of a representative democracy. Candidates for office can position themselves based on party platforms on the horizontal axis from left (liberal) to right (conservative). The vertical axis measures the number of voters that identify with that political position. One additional assumption is that the distribution of voters is normal across the population so that the most voters exist at the median (M).

The median voter model in this context was first presented by Harold Hotelling (1929). Hotelling argued that businesses and politicians, in a two-party system, would locate at the median to attract the most customers and voters. It was Anthony Downs (1957) who first suggested that political parties are strategic in their adopting of policies. According to Downs, "Parties formulate policies in order to win elections, rather than win elections in order to formulate policies" (p. 28). Politicians, as rationally self-interested actors, are concerned with winning elections and will therefore run on policies that will appeal to the decisive median voter. In Figure 23.2, candidate $L_1$ and candidate $R_1$ will both move toward M, where the greatest numbers of voters reside, in order to maximize their vote shares. Thus, candidates want to present policies that appear as centrist or "middle of the road" as possible. This will allow the candidate to attract all the votes to the left of his or her position for $L_1$ (or right for his or her position for $R_1$) plus M—and possibly voters to the right of M if he or she is closer to M than $R_1$ (or left of M if he or she is closer to M than $L_1$). One conclusion from this application is that extreme candidates to

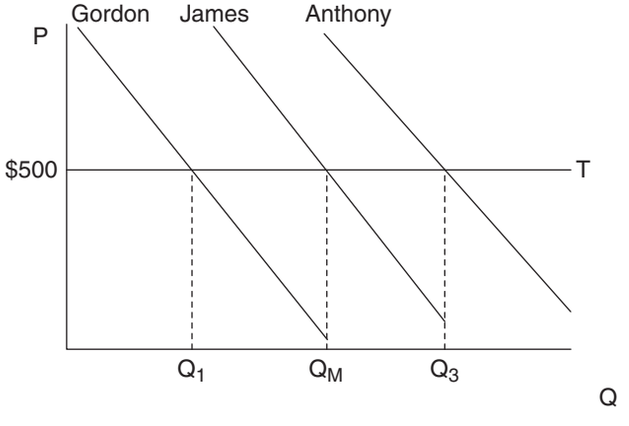


**Figure 23.1** Median Voter Model for a Committee

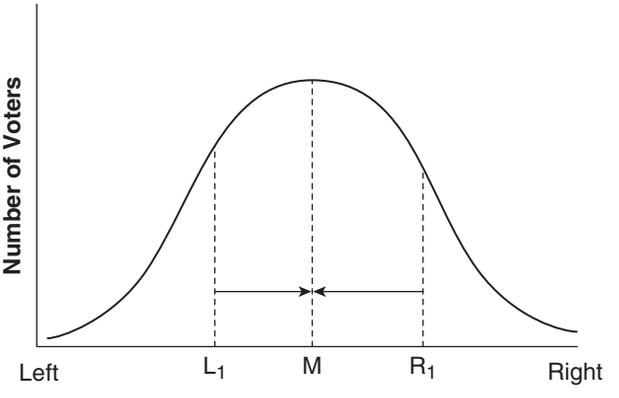


**Figure 23.2** Median Voter Model in a Representative Democracy

either the right or left cannot win elections dominated by a two-party system.

*Agenda Manipulation*

In the presence of voting cycles, the agenda setter has significant power over the final outcome. With a committee structure, in the presence of cycles, the rules that dictate how the agenda will be set matter in determining the end result. An agenda setter, a committee chair, for example, who desires to have his or her preference be the outcome, can structure the votes to achieve this end. Following the example from Table 23.1, suppose that committee members are all voting sincerely for their preferences. Suppose that Gordon is the agenda setter and prefers Option A to B and C. He would pair Option B with C, knowing that B would win, and he would then propose Option A versus Option B. Option A would defeat Option B, and at that point, the agenda setter could suspend all further voting, and the result would be Option A. Joseph Harrington (1990) argues that even when the agenda setter is selected at random, in the presence of voting cycles, that individual has the advantage of acquiring his or her desired outcome over the other members of the committee. Public choice scholars have noticed that despite the focus on voting cycles, there appears to be stability in congressional committee voting (Tullock, 1981). Committees rely on a particular set of rules that allows an agenda setter to avoid these cycles.

*Logrolling*

Economists argue that voluntary exchange is always mutually beneficial, presumably without imposing costs on others. When an opportunity exits for parties to gain from exchange, it is in an individual's rational self-interest to engage in that trade. Political decision making is no exception. One reason that voting may appear stable in the context of cyclical voting is the ability of politicians to exchange votes. Casting a vote does not indicate any intensity of preference. The congressperson from Michigan's vote counts the same as the congressperson from South Carolina's vote. However, the intensity of the preference associated with a piece of legislation may be greater for one congressperson than for another. While the actual buying or selling of votes is illegal, this process has been occurring as an informal practice since the beginning of legislatures. Vote trading or *logrolling* can be described as a "you scratch my back; I will scratch yours" process. Logrolling helps to explain why redistributive polices that benefit only a small segment of the population can be passed, and can continue to be passed, with regularity.

Suppose a congressperson from Michigan would like to pass a tariff that would protect the domestic automobile industry. This tariff will benefit the citizens of Michigan who work for that industry while citizens across the country bear the cost of the tariff. Why would a congressperson from South Carolina support this legislation if his or her constituents would not directly benefit from this policy? If the congressperson from South Carolina would like to receive votes for an infrastructure project, such as a bridge, he or she may vote for the tariff in exchange for the Michigan congressperson's vote. The citizens of Michigan and the other states are unlikely to directly benefit from the building of a bridge in South Carolina in the same way that most citizens will not benefit from the tariff. These types of redistributive policies that benefit a specific congressperson's district or state are known as *pork barrel legislation.*

One can argue that logrolling, especially in the case of pork barrel legislation, is simply tyranny of the majority, because these types of policies benefit one group of citizens at the expense of another. Overall, although trade occurs, there is no increase in net benefit for society, but the congresspeople from Michigan and South Carolina and their constituents may be better off. In the public sector, these benefits come at the expense of others. One conclusion, then, is that majority voting where logrolling occurs will always increase government spending on redistributive policies (Tullock, 1959).

## Public Choice Applied to Representative Democracy

Thus far, the focus has been on direct democracy, where citizens or committees directly vote for some policy, but what about voting for representatives? In the United States, most citizens do not directly vote to pass legislation, although that is possible in the cases of referendums. Rather, they vote for politicians, who will in turn represent them at some level of government.

### Paradox of Voting

Unlike private markets, where an equilibrium outcome exhausts the gains from trade, the costs and benefits associated with voting provide a different outcome. Voting for a representative presents a scenario where the costs outweigh the benefits. Economists argue that individuals will not participate in activities that will not provide a net benefit. If voting does not provide a net benefit, no one should vote, and yet people do vote. This paradox of voting can be better understood by further examining the theory of the rational voter.

*Rational Voter Hypothesis: Instrumental Voting*

The rational voter hypothesis was first developed by Downs (1957) and then further developed by Tullock (1967a) and William Riker and Peter Ordeshook (1968). This topic has generated scores of articles by public choice scholars and continues to be a topic of interest

(see Aldrich, 1993). Here, the simple economic application of cost-benefit analysis is demonstrated. Instrumental voting is simply the voter weighing the expected benefits and costs of voting. The expected benefit of a voter's preferred candidate's win must be greater than the costs associated with voting. For an individual to vote, conditions must be such that $PB - C > 0$. $B$ is the expected benefit that a voter receives if his or her candidate wins the election. $P$ is the probability that one's vote is decisive in determining whether the desired candidate wins the election. In its simplest calculation, one can think of $P$ as $1/N$, where $N$ is the total number of voters participating in the election. Thus, as $N$ gets large, the probability of being the decisive vote for one's candidate becomes infinitesimally small. Dennis Mueller (2003, pp. 304–305) provides a more detailed calculation of $P$. Thus, the probability of one's being the decisive vote for one's preferred candidate is $PB$. $C$ is the cost of voting: the opportunity cost of one's time to actually go the poll and the time one may spend learning about the candidates. Given that the probability of a voter being decisive is perceived as nearly zero (in large elections), even a small cost of voting should keep the rational voter away from the polls.

To complicate matters, if Down's hypothesis regarding candidates' behavior and the median voter is correct, then candidates will quickly move to the center to capture the median voter and win the election. However, if both candidates position themselves at the median, how does the voter perceive a difference between the two candidates? As the candidates move to the middle, the benefit, $B$, one sees for voting for candidate $L_1$ over $R_1$ (to use the notation from Figure 23.2) becomes smaller. Thus, $PB$ may be very small even in cases where $P$ is not infinitesimally small. According to this logic, no one should ever vote, and yet millions of people turn out to vote in elections, hence the paradox. In an attempt to rescue the rational voter hypothesis, Tullock (1967a) and Riker and Ordeshook (1968) included an additional component in the voter's calculation. They suggested that the calculation is $PB + D - C > 0$, where $D$ is a psychic benefit from voting. This psychic benefit can be perceived as a desire to vote or to fulfill one's civic duty. Therefore, if a voter participates in an election, then $D$ must be larger than $C$.

### *Rational Voter Hypothesis: Expressive Voting*

A further development of the rational voter hypothesis is expressive voting. The expressive voting hypothesis picks up where the instrumental hypothesis leaves off. The term $D$, or the psychic benefit one receives from voting, now includes the utility that one receives from being able to express one's preference. Geoffrey Brennan and Buchanan (1984) and Brennan and Loren Lomansky (1993) took a slightly different view of expressive voting, suggesting that voters are still rational in turning out to vote—but not because their votes are decisive. In fact, the hypothesis now assumes that voters understand that they will not cast a decisive vote but instead find value in being able to express their preferences for particular candidates, even if they know that their votes will not change the outcome. Expressive voting has been likened to cheering for one's team at a sporting event. One knows that cheering louder will not change the outcome of the game, but the fan wants to express his or her support for the team.

## Rent Seeking

Perhaps Tullock's (1967b) greatest contribution to the field of public choice is the concept of rent seeking, a term coined by Anne Krueger (1974). In a market economy, economists argue that firms seek profits. These profits (total revenue – total economic cost) provide necessary signals to entrepreneurs regarding how to direct resources. Industries that are earning profits will see increases in resources, and those that are experiencing losses will see resources leave the industry. The profit and loss signals are necessary for the existence of a well-functioning market economy. In all market activities, a business can profit only if it is in fact providing a good or service that is valued by consumers.

Rent seeking is the process of businesses, industries, or special interest groups attempting to gain profit through the political process. Politicians can supply policy that provides an economic advantage to a particular group, in the form of granting them monopoly power or regulation that creates barriers to entry. The policy would benefit this group at the expense of the existing or potential competitors. These rents from the creating, increasing, or maintaining of a group's monopoly power involve payoffs large enough that groups are willing to expend effort to acquire them. More important, the process of rent seeking, unlike profit seeking, results in costly and inefficient allocation of resources.

Suppose Figure 23.3 represents the market for cosmetologists (hair dressers). D is the demand for these services, and $S_C$ is the supply in a competitive market. A simplifying assumption is made that the marginal cost of serving one more customer is constant; hence, $S_C$ is horizontal. If this market operates competitively, the consumers will pay a price of $P_C$ and be supplied a quantity of $Q_C$. Now suppose that the existing members of the cosmetology industry hire lobbyists to persuade government officials to impose new regulations on the industry that create higher barriers to entry, giving existing members greater monopoly power. After the regulation is imposed, the costs of becoming a cosmetologist increase, causing the supply curve to shift from $S_C$ to $S_M$. Consumers now pay the higher price of $P_M$, and the cosmetologists will reduce the quantity to $Q_M$.

In a traditional monopoly analysis, the area labeled A would be known as the welfare transfer. The suppliers

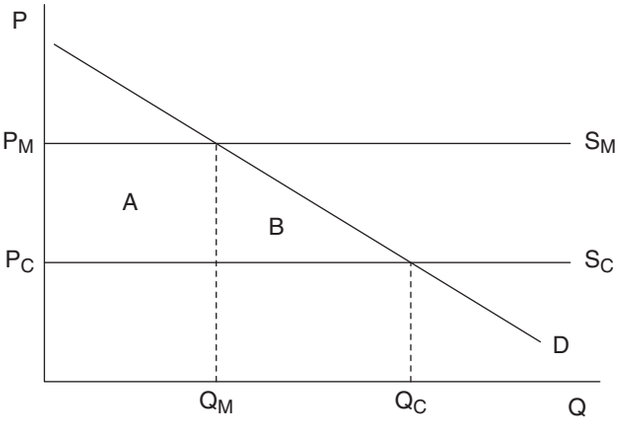


**Figure 23.3** Rent Seeking in the Cosmetology Market

gain Area A at the expense of the consumers in this market. These are the rents that the cosmetology industry is seeking through the political process. Area B is the deadweight loss, the loss of consumer and producer surplus. However, Tullock (1967b) argued that both Areas B and A are deadweight losses. Area A is the rents that can be extracted by the group through the use of government power. In this example, the cosmetology industry is willing to pay up to the amount of A to receive these rents. Real resources must be used to acquire these rents. The industry would hire lobbyists, lawyers, and consultants to meet with government officials. It may engage in a media campaign to justify the new regulations. In addition, there are other groups, perhaps consumers or a competing interest group, that may spend resources in an attempt to prevent this regulation. These are costs above and beyond what may be gained by the cosmetology industry. In the end, Area B is a deadweight loss because real production is lost from $Q_C$ to $Q_M$. Area A is a deadweight loss, or wasteful spending, because real resources are used to acquire these rents rather than to produce more real goods and services, and this creates regulation that promotes inefficiency in the market.

### *Rational Ignorance*

If economists (and presumably government officials) realize that rent seeking is wasteful, why does it occur so often? To answer this question, we return to the supply and demand for public policy. Politicians can supply policy to different groups of voters: large, unorganized groups such as the working poor, the middle class, or college students; smaller, well-organized groups such as unions; industry-specific groups such as automobile manufacturers; or groups organized by occupation such as farmers. The unorganized groups are generally classifications of citizens that have only their individual votes. The organized or special interest groups provide contributions and often a large bloc of votes to a politician.

We know from our discussion of the rational voter hypothesis that the benefits from voting are extremely small, and the cost of voting may outweigh these benefits. More important, the costs of voting do not simply include the costs of voting on the day of the election but also the costs of knowing who or what is on the ballot. Being a well-informed voter is extremely costly. Voters have to research candidates' records and policy positions, which are often not fully provided by the media or in the candidate's literature. Citizens have to be aware of the goings-on within in their legislative body. The problem is that being a well-informed voter does not change the probability or benefit of voting. Well-intentioned and informed voters might participate only to see none of their proposals or candidates win. Thus, there is little benefit to being a well-informed voter. Therefore, the self-interested voter is rationally ignorant.

Voters focus on the few policies that matter to them and become aware of only whether the politician supports their view. It is rational for voters to remain ignorant of most issues and incur only the costs associated with the issues that are most important to them. A consumer shopping for a new car will not test-drive every make and model before purchasing the car. At some point, the additional information that can be gained from driving another car will be less than the cost associated with that test-drive. The same idea applies to voters: Beyond a certain point, trying to gather more information about a candidate will prove too costly for the benefit, and voters choose to be ignorant.

It is this rational ignorance of voters that allows wasteful rent-seeking policies to occur. Most citizens remain unaware of the regulations surrounding a particular industry and the motivations for imposing them. It is also important to realize that well-organized special interest groups have the incentive to be well informed. The benefits to the special interests are large, relative to the dispersed costs imposed on the taxpayers.

## Legislatures, Bureaucracies, and Special Interest Groups

If politicians want to maximize votes to win elections, they can appeal to rationally ignorant voters in ways that convince the citizens to vote for them. For example, a politician campaigning on a college campus may speak on issues of affordable higher education. Rationally ignorant voters now know that this candidate will promote a policy that will benefit college students. These students may not seek any additional information on the candidate and may proceed to vote for him or her. Building on the concepts developed in this chapter, we can see that legislators will supply public policy to special interest groups in addition to broad groups of voters. Legislators using logrolling and relying on the rational ignorance of voters will allow special interest groups to rent seek in an effort to win or remain in office.

Special interest groups are typically organized based on industry, occupation, or a political cause. Mancur Olson (1965) argued that when these groups form to promote the interests of the group, they take on the characteristics of public goods, which means they suffer from the free-rider problem. Therefore, he argued that these groups will be more effective when they are relatively small, and if they are large, they require what Olson referred to as *selective incentives* to minimize the free-rider problem. An example of a selective incentive is a union's requiring that members have dues deducted from their wages so that members cannot receive the benefits of the union without incurring the costs. Public choice economists have made two arguments regarding why special interest groups contribute to elections: (1) in an effort to (re)elect politicians and (2) to influence the way a politician votes on a public policy relevant to the special interest group.

Political action committees (PACs) are the legal or formal method by which special interest groups make contributions to legislators. The empirical literature on the influence of PACs on elections and legislation is numerous. James Kau and Paul Rubin (1982, 1993), Michael Munger (1989), Kevin Grier and Munger (1991), and Thomas Stratmann (1991, 1992, 1995, 1996, 1998) find evidence that PACs give to a legislator, with the hope of influencing the legislator's votes and not merely to improve his or her chance to win office. For a brief review of this literature, see Mueller (2003). Calcagno and Jackson (1998, 2008) argue that PAC spending increases roll call voting, the voting on legislation, for members of the U.S. Senate and the House of Representatives. Thus, public choice economists argue that special interest groups have more influence on political decision making for three basic reasons: (1) Politicians want to win elections, and special interest groups (PACs) can offer money and votes to improve the odds of winning; (2) voters are rationally ignorant of the policies that special interest groups are lobbying for or against; and (3) these policies offer concentrated benefits to special interest groups and widespread costs on individual voters.

Bureaucrats run the government on a day-to-day basis and include everyone from the clerk at the Department of Motor Vehicles to the head of the Food and Drug Administration. These individuals are primarily appointed and are not subject to the same accountability as politicians. These individuals cannot be voted out of office and many remain in these positions long after the politicians that appointed them have left office. William Niskanen (1971) was one of the first economists to examine bureaucracies from a public choice perspective. If bureaucrats, like everyone else, are rationally self-interested, what is it that they are maximizing? Niskanen argued that bureaucrats are budget maximizers. Larger budgets signal greater political power, prestige, and job security. Bureaucracies are difficult to monitor. This allows the bureaucrat to extract these nonpecuniary benefits. The model of bureaucracy is based on the monopoly model. The bureau is the only one producing a particular good or service for the public. Bureaucracies present to congress and the citizens an all-or-nothing demand curve. That means they can either produce the goods or services for a specific budget or not at all. This perception is in part due to the difficulty of monitoring bureaus. Beyond that, the output of bureaucracies is often difficult to measure, making it easier to present to the public a level of output that requires a specific budget.

Unlike the private sector, where managers have the incentive to improve production and are rewarded for their efficiency by higher profits, bureaucrats do not receive any such reward. Thus, the incentive to improve efficiency is not part of the bureaucracy's institutional structure. Bureaucrats will seek larger and larger budgets under the guise of satisfying citizens. In addition, bureaus can appeal not only to satisfying the citizenry as a whole but also often provide a good or service that is specific to a special interest group. In an effort to expand the bureau's patronage, bureaucrats make the special interest group aware of the goods or services they provide that benefit the group. Providing these goods or services encourages special interest groups to pursue rent-seeking behavior—that is, more benefits for the special interest group. Providing more goods or services allows bureaus to expand in size, scope, and budget.

### *Government as Leviathan*

If politicians are maximizing votes, bureaucrats are maximizing budgets, special interest groups are dominating the political process, and voters are rationally ignorant, then the logical conclusion is the theory of government as Leviathan. Rather than a benevolent government, which seeks to aid the public's interest, the theory of Leviathan suggests government is a monopolist with the sole interest of maximizing revenue. Government officials use cooperation—or whatever means they are constitutionally permitted—to generate as much revenue as citizens will allow. This view of government is at odds with traditional public finance, which suggests that there are optimal levels at which to tax citizens and that government officials are constrained by their accountability to citizens. If citizens (voters) are uninformed, government officials with the power to tax, spend, and create money are uncontrolled. Even with constitutional limits, rationally ignorant voters will not understand the amounts they are being taxed, the size of deficits and debt, and the impact of expanding money supplies. When government exceeds its constitutional limits, citizens may realize that they must monitor government more closely. However, these attempts to constrain government are often short lived. Voters attempted to constrain government with tax revolts in the 1970s and with the Contract with America (Mueller, 2003) in the early 1990s. Brennan and Buchanan (1980) provide a complete analysis of the theory of government as Leviathan.

*Political Business Cycle*

Although public choice is primarily a microeconomic field, it examines the macroeconomy through the prism of business cycles. Business cycle theories can be broadly categorized by their causes: (a) large declines in aggregate demand, (b) expansionary monetary policy, and (c) technological shocks to the economy. While business cycle theories may suggest that government policy may play a role in the creation of the business cycle, public choice scholars take this view one step further. Business cycles are the result of government policy, primarily through monetary expansion, and these cycles coincide with election cycles. Political business cycle economists argue that politicians will attempt to improve economic conditions prior to elections so as to win reelection. Focusing on a simple trade-off between unemployment and inflation, politicians will pursue policies that will reduce unemployment and generate positive economic conditions. Monetary expansion and fiscal policy are used to reduce unemployment. Only after the election does the inflationary effect of the policy emerge, and the boom becomes a bust. At this point, the politicians start the cycle again. This opportunistic political business cycle in part relies on the rationally ignorant voter, but one wonders if voters are fooled over and over again by such policies. An alternative explanation is that voters are selecting politicians that they believe will benefit them. If unemployment is a concern for a voter, he or she may vote for a politician who reduces unemployment because it will specifically benefit the voter.

According to Jac Heckelman (2001), the first term of the Nixon administration provides anecdotal evidence of a political business cycle, and Heckelman suggested that it was Nixon's administration that inspired early theoretical models:

> Keller and May (1984) present a case study of the policy cycle driven by Nixon from 1969–1972, summarizing his use of contractionary monetary and fiscal policy in the first two years, followed by wage and price controls in mid-1971, and finally rapid fiscal expansion and high growth in late 1971 and 1972. (p. 1)

Similarly, Jim Couch and William Shughart (1998) find evidence that New Deal spending was higher in swing states than in states with greater economic need, suggesting that the spending was motivated by Roosevelt's desire to be reelected. The empirical evidence of an opportunistic political business cycle has mixed results in the economics literature. In a recent study, Grier (2008) found strong support for the political business cycle. He found evidence that the real gross national product (GDP) rose and fell between the years 1961 and 2004, in conjunction with presidential election cycles.

## Normative Public Choice

### The Importance of Institutions

Like most economics, public choice is primarily positive economic analysis, which means it is used to explain what is or what will happen. However, because public choice examines the world of political decision making and these decisions affect factors such as economic growth, taxation, and individuals' standards of living, it seems appropriate that it ventures into the realm of normative analysis, or the way things should or ought to work. The use of normative economics in public choice has focused on the role of institutions. It is not enough to understand why voters may chose a particular candidate under majority rule, but one must ask whether majority rule is the best institution to select elected officials. Gwartney and Wagner (1988) argued that if individual actors are going to generate outcomes that benefit society, then

> success rests upon our ability to develop and institute sound rules and procedures rather than on our ability to elect "better" people to political office. Unless we get the rules right, the political process will continue to be characterized by special interest legislation, bureaucratic inefficiency, and the waste of rent-seeking. (p. 25)

*Constitutional Economics*

The normative aspect of public choice can be seen as the overlap with social choice theory. One major normative analysis that emerges from public choice theory is constitutional economics. As in social choice theory, the fundamental question addresses the best way for a society to be organized so as to maximize the welfare of its citizens. It is precisely this question that Buchanan and Tullock (1962) tried to answer in *The Calculus of Consent.* Constitutional economics can be thought of as both a positive and a normative analysis. In normative analysis, the argument for a constitution is similar to other social choice theories that address the concept of a social contract. Public choice theory starts with the first stage, a convention being called, to establish a constitution under uncertainty. The difference in these theories is the degree of uncertainty for the actors in this first stage of decision making. Buchanan and Tullock argue that in this first stage, collective decisions are really individual decisions in which each individual follows his or her own self-interest.

The constitutional convention requires unanimity to have a successful outcome. If everyone agrees to the constitution, everyone is agreeing to abide by the rules, even though some of those rules may ban or restrict their behaviors in the future. It is in this way that the uncertainty comes into play. The actors know that certain actions will reduce the utility of individuals in the future, but they do not know at that point whether they will be one of the individuals affected. A first-stage decision may be the voting rule used in future

elections: simple majority rule, a two-thirds majority, or a three-fourths majority. By having unanimity in the first stage, everyone is agreeing to that rule without knowing whether they will be in the majority in the future. It is not until the second stage when individuals are actually voting on policies and candidates that they will know to what degree they will be affected by the first-stage decisions.

One contention in this theory is whether constitutions are contracts or conventions. The contract theory views constitutions as contracts agreed on and actually signed by the originators to signify their commitment. The convention view suggests that constitutions are devices by which social order and collective decisions are made. The convention view treats constitutions as self-enforcing, because all parties understand the convention or device by which decisions will be made and chose a convention in the first stage that maximizes social welfare. Thus, the second-stage decisions are easily agreed on, and no one has an incentive to violate the convention. The constitution-as-contract view requires an outside enforcement mechanism, such as a police force or a court system, to ensure that individuals do not violate the contract. Both views recognize that at the first stage, the rules need to be constructed to create an incentive structure consistent with individuals' rational self-interest. Otherwise, in the second stage, individuals will generate outcomes that will not maximize social welfare. Constitutions possess characteristics of both contract and convention. Although the self-enforcing feature of conventions is desirable, one potential problem is that constitutions can evolve over time, and these decisions may be out of the citizens' control. When a constitution is a contract, it is harder to change, but because the government enforces it, conflicts between the state and the citizen can arise.

### *Anarchy*

It is precisely the idea of the state as a neutral party that has some public choice scholars asking whether constitutions are the right means of organizing society. As has been noted throughout this chapter, individuals, in the absence of sound institutions, will create outcomes that are inefficient. Instead of a constitutionally limited state, some economists argue that the logical conclusion of a public choice perspective is anarchy. Beginning in the 1970s, public choice economists began to examine the issue of anarchy. The initial results were pessimistic views that a state is still necessary to prevent individuals from plundering their neighbors. However, a younger generation of public choice economists has a more optimistic view (Stringham, 2005). The absence of government does not mean the absence of rules or laws but rather the form that they take. Private institutions, through voluntary exchange and contracts, will create a set of rules that individuals agree to abide by. This research is being revisited, and this question is being asked: Why not consider anarchy in the realm of possible choices?

Peter Boettke (2005) pointed out that public choice economics presents a puzzle. He argued that Buchanan suggested there are three potential functions for government: (1) protective, in which the state is protector of private property rights; (2) productive, in which the state improves inefficiency where warranted; and (3) redistributive. The first two features, many economists will argue, are desirable, but the third, the redistributive power of the state, is the one that exists but should be avoided. How do you create a state that will generate the first two and not the third? Boettke suggested that there is a fine line between opportunism and cooperation in individuals' self-interested behavior. Does this require the state's intervention, or can markets solve this puzzle? This issue provides an interesting area for future research in public choice.

## Conclusion and the Future of Public Choice

It has been said the framers of the U.S. Constitution were the original public choice economists. Somewhere between the eighteenth century and the first half of the twentieth century, this view of politics and institutions changed. Beginning in the 1940s and taking hold in the 1960s, public choice economics emerged to return us to this original vision. Using economic tools to examine political choices provided new insights into the field of public finance. The dichotomy of private decision making as self-interested and public decision making as altruistic was confronted with a single vision of economic humanity. Individuals, whether they are buying a car or deciding for which congressperson to vote, are now seen as behaving rationally self-interested.

Over the last half-century, the field of public choice has examined every aspect of political decision making from voting to politicians' behavior and from campaign financing to the running of bureaucracies. The field of public choice has questioned whether political behavior drives our macroeconomy, debated the proper origins of the state, and asked whether the state can be contained and whether the state should even exist. Public choice economics has been integrated into the field of public finance and neoclassical economics to such an extent that not treating both private and public actors as rationally self-interested may be viewed as an incomplete analysis, perhaps even naive.

Public choice has tackled many of these issues, empirically finding strong evidence of rationally ignorant voters, rent seeking, vote-maximizing politicians, and the existence of an opportunistic political business cycle. It has ventured in the normative arena, asking what type of institutional arrangements will generate the greatest prosperity and maximize social welfare.

What has emerged from this field is an understanding that if market failure exists, so does government failure. Public choice economics presents a means by which market-based outcomes can be compared to public-sector

outcomes and economists can decide which institutional arrangement, even if flawed, will likely lead to greater economic efficiency (Gwartney & Wagner, 1988).

Public choice theory continues to examine these issues, both empirically and from a normative perspective. Public choice theory continues to make contributions to the field of economics, analyzing every type of collective decision. With the government taking a larger and larger role in the economy, public choice will continue to prove a useful field of analysis.

## References and Further Readings


Aldrich, J. H. (1993). Rational choice and turnout. *American Journal of Political Science, 37,* 246–278.

Arrow, K. J. (1951). *Social choice and individual values.* New York: John Wiley.

Arrow, K. J., Sen, A. K., & Suzumura, K. (Eds.). (2002). *Handbook of social choice and welfare* (Vol. 1). New York: Elsevier North-Holland.

Black, D. (1948). On the rationale of group decision-making. *Journal of Political Economy, 56,* 23–34.

Boettke, P. J. (2005). Anarchism as a progressive research program in political economy. In E. Stringham (Ed.), *Anarchy, state, and public choice* (pp. 206–219). Northampton, MA: Edward Elgar.

Brennan, G., & Buchanan, J. M. (1980). *The power to tax: Analytical foundations of a fiscal constitution.* Cambridge, MA: Cambridge University Press.

Brennan, G., & Buchanan, J. M. (1984). Voter choice: Evaluating political alternatives. *American Behavioral Scientist, 28,* 185–201.

Brennan, G., & Lomansky, L. (1993). *Democracy and decision.* Cambridge, MA: Cambridge University Press.

Buchanan, J. M. (1949). The pure theory of government finance: A suggested approach. *Journal of Political Economy, 57,* 496–506.

Buchanan, J. M. (1954). Individual choice in voting and the market. *Journal of Political Economy, 62,* 334–343.

Buchanan, J. M. (2003). *Public choice: The origins and development of a research program.* Fairfax, VA: George Mason University, Center for Study of Public Choice.

Buchanan, J. M., & Musgrave, R. A. (1999). *Public finance and public choice: Two contrasting views of the state.* Cambridge: MIT Press.

Buchanan, J. M., & Tullock, G. (1962). *The calculus of consent: Logical foundations of constitutional democracy.* Ann Arbor: University of Michigan Press.

Calcagno, P. T., & Jackson, J. D. (1998). Political action committee spending and senate roll call voting. *Public Choice, 97,* 569–585.

Calcagno, P. T., & Jackson, J. D. (2008). Political action committee spending and roll call voting in the U.S. House of Representatives: An empirical extension. *Economics Bulletin, 4,* 1–11.

Couch, J. F., & Shughart, W. F. (1998). *The political economy of the New Deal.* Cheltenham, UK: Edward Elgar.

Downs, A. (1957). *An economic theory of democracy.* New York: Harper & Row.

Grier, K. B. (2008). Presidential election and real GDP growth, 1961–2004. *Public Choice, 135,* 337–352.

Grier, K. B., & Munger, M. C. (1991). Committee assignments, constituent preferences, and campaign contributions. *Economic Inquiry, 29,* 24–43.

Gwartney, J., & Wagner, R. E. (1988). The public choice revolution. *Intercollegiate Review, 23,* 17–26.

Harrington, J. E. (1990). The power of the proposal maker in a model of endogenous agenda formation. *Public Choice, 64,* 1–20.

Heckelman, J. (2001). Historical political business cycles in the United States. In R. Whaples (Ed.), *EH.Net encyclopedia.* Available at http://eh.net/encyclopedia/article/heckelman.political.business.cycles

Holcombe, R. (2005). *Public sector economics: The role of government in the American economy.* Upper Saddle River, NJ: Prentice Hall.

Hotelling, H. (1929). Stability in competition. *Economic Journal, 39,* 41–57.

Kau, J. B., & Rubin, P. H. (1982). *Congressmen, constituents, and contributors.* Boston: Martinus Nijhoff.

Kau, J. B., & Rubin, P. H. (1993). Ideology, voting and shirking. *Public Choice, 76,* 151–172.

Keller, R. R., & May, A. M. (1984). The presidential political business cycle of 1972. *Journal of Economic History, 44,* 265–271.

Krueger, A. O. (1974). The political economy of the rent-seeking society. *American Economic Review, 64,* 291–303.

Mueller, D. C. (2003). *Public choice 3.* New York: Cambridge University Press.

Munger, M. C. (1989). A simple test to the thesis that committee jurisdictions shape corporate PAC contributions. *Public Choice, 62,* 181–186.

Niskanen, W. A. (1971). *Bureaucracy and representative government.* Chicago: Aldine-Atherton.

Olson, M. (1965). *The logic of collective action.* Cambridge, MA: Harvard University Press.

Riker, W. H., & Ordeshook, P. (1968). A theory of the calculus of voting. *American Political Science Review, 62,* 25–42.

Stratmann, T. (1991). What do campaign contributions buy? Causal effects of money and votes. *Southern Economic Journal, 57,* 606–620.

Stratmann, T. (1992). Are contributors rational? Untangling strategies of political action committees. *Journal of Political Economy, 100,* 647–664.

Stratmann, T. (1995). Campaign contributions and congressional voting: Does the timing of contributions matter? *Review of Economics and Statistics, 77,* 127–136.

Stratmann, T. (1996). How reelection constituencies matter: Evidence from political action contributions and congressional voting. *Journal of Law and Economics, 39,* 603–635.

Stratmann, T. (1998). The market for congressional votes: Is timing of contributions everything? *Journal of Law and Economics, 41,* 85–113.

Stringham, E. (Ed.). (2005). *Anarchy, state, and public choice.* Northampton, MA: Edward Elgar.

Tollison, R. D. (1982). Rent seeking: A survey. *Kyklos, 35,* 575–602.

Tullock, G. (1959). Some problems of majority voting. *Journal of Political Economy, 67,* 571–579.

Tullock, G. (1967a). *Toward a mathematics of politics.* Ann Arbor: University of Michigan Press.

Tullock, G. (1967b). The welfare costs of tariffs, monopolies, and theft. *Western Economic Journal, 9,* 224–232.

Tullock, G. (1981). Why so much stability. *Public Choice, 37,* 189–202.

# 24

# Taxes Versus Standards

## *Policies for the Reduction of Gasoline Consumption*

Sarah E. West
*Wesleyan College*

Fossil fuel consumption and greenhouse gas emissions from the personal transportation sector promise to be key challenges facing the next generation of policy makers. In 1999, transportation became the largest end-use producer of carbon dioxide from fossil fuels in the United States. Gasoline-powered vehicles are responsible for about 60% of the carbon from the transportation sector. They also emit the majority of carbon monoxide, as well as substantial proportions of oxides of nitrogen and volatile organic compounds (Davis, 2004).

The transportation sector became the largest producer of carbon dioxide for two reasons. First, with the exception of 2007 and 2008, when gas prices reached new highs and the U.S. economy slumped into recession, households increased the number of miles they drove (Federal Highway Administration, 2009). Second, the vehicle fleet has become less fuel efficient so that in 2005, households on average used more gasoline per mile driven than they used per mile in 1987 (Davis, 2007). This is a stunning statistic, given that, overall, the U.S. economy uses far less energy per unit of output now than it did 30 years ago.

Why did the transportation sector become less fuel efficient? After all, passenger cars sold now get about 7 or 8 more miles per gallon, on average, than those sold in 1978. And light-duty trucks, the category that includes pickups, vans, and sport-utility vehicles (SUVs), experienced an increase in fuel efficiency of about 3 miles per gallon since 1978. But the market share of light-duty trucks, which are less fuel efficient than cars, increased dramatically over that time period, rising from less than one third of vehicles purchased in 1988 to more than half by 2000. Increases in SUV purchases were primarily responsible for the increase. The increase in the number of these less fuel-efficient vehicles, relative to the number of passenger cars, brought down the overall fuel economy of the U.S. vehicle fleet.

At the same time, problems related to traffic congestion and dependence on foreign oil intensified, and awareness of the likely impacts of climate change increased. Policy makers have therefore directed their attention toward policies for the reduction of gasoline consumption.

This chapter summarizes the current state of economic knowledge on whether government intervention in gasoline and fuel economy markets is justified, and if so, what form policies should take if they are to efficiently and effectively reduce gasoline consumption. It begins by defining market failure due to the presence of externalities and by describing the kinds of policies generally used to address such market failures in the United States. It then quantifies the external costs associated with gasoline consumption, concluding that the presence of a number of externalities in gasoline markets justifies government intervention in those markets. Following discussion of the relative efficiency of various policies that may correct failures in the gasoline market, this chapter considers the possibility that not only do consumers fail to internalize the costs they impose on others by consuming gasoline but they also fail to fully account for the fuel

savings that can be had by buying a more fuel-efficient vehicle. If consumers are fallible in this way, then not only is intervention in gasoline markets justifiable on efficiency grounds, but so is intervention in the market for fuel economy. The same is true if automakers fail to provide a sufficiently broad and varied selection of fuel-efficient vehicles. The chapter concludes with a discussion of distributional considerations of policies for the reduction of gasoline consumption and with suggestions for directions for future research.

## Justifications for and Types of Government Intervention

When consumers decide how much gasoline and the kind of vehicle to buy, or producers decide how much to sell, they weigh the private costs of their activities against the benefits. For example, consumers evaluate the prices of various vehicles, the costs of operating such vehicles, the price of gasoline, and the prices of alternative modes of transport. Producers compare the revenues that they expect to earn with the costs of refining and distributing gasoline or of producing various types of vehicles. Without proper incentives, however, consumers and producers will not include the costs that they impose on the environment and others in their decisions of how much and what to consume and produce. This failure to include these external costs in decisions results in a market failure: Drivers buy and use and producers sell too much gasoline and vehicles that are too inefficient. When market failure occurs, usually a case can be made for government intervention (West & Wolverton, 2005).

Governments have typically used both command-and-control (CAC) regulation and market-based incentives to resolve market failures. Prior to 1990, virtually every environmental regulation in the United States and elsewhere took the form of CAC regulations, and they are still commonly used. These regulations are so named because they command that emissions be controlled to meet a given minimum or maximum standard. As such, they tend to be either technology based or performance based.

Technology-based regulations mandate the control technology or production process that producers must use to meet the emissions or fuel economy standard set by the government. Such a standard might take the form of a requirement that a particular kind of vehicle, an electric vehicle for example, be adopted by all state and local governments. One problem with this type of CAC regulation is that it applies a one-size-fits-all policy to producers and consumers that may face very different costs of reducing gasoline consumption. Thus, while gasoline is reduced to the desired level, it is accomplished at a higher cost than might have occurred if firms and consumers were allowed to determine the most cost-effective means for meeting the standard. Alternatively, if more flexible policies were used, more gasoline reduction or higher environmental quality and fuel efficiency could be achieved at the same cost. Technology-based CAC policies do not encourage vehicle manufacturers or consumers to find new and innovative fuel-saving strategies, nor do they provide an incentive to reduce consumption beyond the set level.

Performance-based regulations are more flexible CAC policies; they mandate that a standard be reached but allow producers and consumers to choose the method by which to meet the standard. Still, once they have reached the level specified by the standard, producers and consumers face little incentive to increase fuel economy or reduce gasoline consumption any further.

Market-based policies, on the other hand, are regulations that encourage behavior through price signals rather than through explicit instructions or mandates (Stavins, 2000). Many market-based instruments function as follows: A firm or a consumer faces a potential penalty in the form of a tax or permit price per unit of gasoline, emissions, or fuel inefficiency. The firm or consumer can choose to pay for existing gasoline or low fuel economy via the tax or permit, or can instead reduce gasoline consumption or increase fuel economy to avoid paying the penalty. Other market-based policies might subsidize fuel economy. For example, the federal government subsidizes hybrid vehicles until manufacturers reach a specific production volume, and several states provide similar incentives.

Market-based policies give consumers and automakers more flexibility than most CAC policies. First, the method for reducing gasoline consumption is not specified, giving drivers or manufacturers with heterogeneous costs the flexibility to use the least costly fuel-saving method. Second, because market-based incentives force producers and consumers to pay taxes, buy permits, or forgo subsidies when they sell or drive fuel-inefficient vehicles, they provide an always-present incentive to take measures that reduce gasoline consumption. Such incentives, therefore, also promote innovation in fuel-saving technologies (for discussion of the effects of incentives on innovation, see, for example, Jaffe & Stavins, 1995; Laffont & Tirole, 1996: Parry, 1998).

Only in a very narrow set of circumstances could such market failures be addressed without government intervention. In a seminal article, Ronald Coase (1960) lists three conditions that must be met for an externality problem to be resolved without government intervention. There must be (1) well-established property rights, (2) a willingness of affected parties to bargain, and (3) a small number of parties affected. Although it might be possible to legislate rights for, say, fuel economy or a stable climate, it is less likely that oil companies, automakers, and those affected by the external costs of driving would be willing to bargain effectively. And as the number of parties affected by climate change includes the global population, Coase's third condition is clearly violated.

## Failures in the Market for Gasoline

Think about the last time you drove or rode in a car. Did you consider the fact that your car pollutes the air with hazardous emissions, adds greenhouse gases to a warming atmosphere, increases congestion and the likelihood of accidents on the road, and exacerbates reliance on oil imported from areas prone to war and dictatorial regimes? More important, if you did think about these external effects of your gasoline consumption and driving, did you drive or ride in a car less than you would have had gasoline consumption been harmless, and did you opt for a more fuel-efficient vehicle? Maybe you did. But do most Americans? Probably not. This means that even if you are paying $4.00 per gallon for gasoline, you are paying too little. That $4.00 covers the costs of oil, refining, distribution, marketing, and state and federal gas taxes. But as we shall see below, the taxes per gallon fall far short of covering external costs.

### Quantifying the External Costs of Gasoline Consumption

The external costs related to gasoline consumption are substantial, which implies that in the absence of government intervention, the gasoline market fails to operate efficiently. Ian Parry, Margaret Walls, and Winston Harrington (2007) divide these costs into two categories: per gallon externalities and per mile externalities.

Per gallon externalities involve costs imposed on others when gasoline is burned, regardless of how many miles were driven when it was burned. Parry et al. (2007) suggest that a reasonable estimate for the costs of greenhouse gases per gallon of gasoline is 6¢, while the estimate of costs due to increased oil dependency is 12¢ per gallon (both in 2007 U.S. dollars).

Of course, uncertainty in the science of climate change and its effects, not to mention uncertainty about the relationship among gasoline consumption, oil dependency, and terrorism, imply that the true magnitude of these costs is itself extremely uncertain. But even if the per gallon external costs of gasoline consumption are much higher than the best guesses in Parry et al. (2007), it is very likely that the per mile external costs swamp per gallon costs. Traffic congestion in the United States causes drivers to lose billions of dollars per year of one of their most valuable resources: time. For this reason, external costs due to congestion are estimated to be a whopping $1.05 per gallon (in 2007 dollars). Costs from increases in accidents due to increases in gasoline consumption when more miles are driven, plus costs from local pollutants such as carbon monoxide and ozone, add up to another $1.05, for a total of $2.10 per gallon in per mile external costs of gasoline consumption.

In addition to the per mile costs discussed by Parry et al. (2007), an increase in miles driven also involves a more subtle but sizable cost. The argument goes like this: Because they face a tax that depends on how many hours they work, workers work too little, which implies that labor markets operate inefficiently. Any policy that further reduces the number of hours worked will exacerbate this inefficiency. It turns out that making either gasoline or miles cheaper, and thereby increasing miles driven, also reduces hours worked. If it costs a worker less to drive, for example, he or she might be more likely to take work off early on a Friday to drive to a weekend retreat. Sarah West and Roberton Williams III (2007) find that because labor markets are so large, small reductions in hours worked that occur when miles driven increase involve potentially large efficiency losses, with the costs of reduced work hours amounting to about a third of the sum of the external costs estimated by Parry et al. (2007).

The magnitude of these external costs, when taken together and when compared to each other, has two important complications. First, existing federal and state gasoline taxes, which sum to an average of 40¢ per gallon, do not cover external costs. Second, because the costs associated with driving a mile far exceed the costs of burning the gasoline in isolation, any policy that increases miles driven is likely to be quite costly in terms of its effects on congestion and accidents.

### Policies to Correct Failures in Gasoline Markets

It seems clear that gasoline markets will be inefficient if left to themselves. What should be done, then, to induce gasoline producers, automakers, and drivers to internalize the costs they impose on others? Any policy that does this successfully must induce consumers to use less gasoline. Mindful of the relatively high per mile costs of congestion and accidents, one can place policies that reduce gasoline consumption into two categories: those that also reduce miles driven and those that increase miles driven.

In the absence of an additional market failure (e.g., a failure in the market for fuel economy, discussed at length below) or political or administrative constraints, a policy that also reduces miles driven—or at least one that does not increase miles driven—will be preferable to policies that reduce gasoline consumption but also increase miles driven.

Policies that increase the price paid per gallon or the price paid per mile can be expected to reduce both gasoline consumption and miles driven. Such policies include a tax per gallon of gasoline, a carbon tax or cap-and-trade program, a direct tax on vehicle-miles driven, or a pay-as-you-drive (PAYD) insurance premium (see Parry et al., 2007, for general discussion and Parry, 2004, 2005, for a more specific treatment of the PAYD insurance premium).

The gasoline tax has the advantage of being easy to administer, and in terms of per gallon externalities, it internalizes costs perfectly. It efficiently encourages drivers to consider their effects on the climate, because the amount

of carbon released by a gallon of gasoline is independent of the manner in which it is combusted. The same is true for the effects on oil dependency. It is important to note that a gasoline price floor, which would set a minimum price for gasoline, would not attain the same efficient outcome as would a gasoline tax. Such a policy, which has been discussed in the U.S. Congress, would likely lead to inefficient gasoline shortages, when binding, and inefficiently low gasoline prices, when not binding.

When faced with a tax on gasoline, drivers will choose the most cost-effective method of reducing consumption. Consumers that live near work may opt to walk, bike, or take public transportation. Others might choose to change air filters more often, to monitor tire pressure more vigilantly, or to reduce highway driving speed. An increased tax on gasoline would also induce consumers to change their vehicle purchasing behavior. They might buy a Toyota Camry with a four-cylinder engine when they would have otherwise purchased one with six cylinders. Or they may replace a less efficient car with an established hybrid, such as a Toyota Prius, or with one of the newly emerging vehicle technologies. Advances in battery technology have now moved electric vehicles off the drawing board. High-end electric vehicles are currently available from Tesla Motors, and General Motors's Volt electric car is scheduled to go on sale in 2010. Low-end, inexpensive golf-cart type vehicles are being upgraded to meet crash protection standards. Plug-in hybrids are being readied for production. Gasoline taxes would increase the demand for these vehicles and reduce the demand for conventional gasoline-powered cars and trucks.

A tax on gasoline would also induce consumers to seek cheaper fuel alternatives for their conventional vehicles. About 49% of the gasoline in the United States is already E10 (gasohol), which is 10% ethanol. Brazil has completely freed itself from foreign oil dependence, with virtually all its cars running on E85, which is 85% ethanol. There are already 6 million "flex-fuel" vehicles in the United States that can run on E85 right now. Used conventional cars can be converted to flex-fuel for about $100 in parts, including replacing the computer chip that controls the air mixture, attaching new fittings to the fuel lines, and replacing the rubber seals with nonrubber seals. Wal-Mart is working out the details with Murphy Oil Company to make E85 available at Sam's Club and Wal-Mart gas stations throughout the United States.

Switching to corn-based E85 is likely to reduce the per gallon costs of oil dependence, but it is also likely to result in higher greenhouse gases and greater health costs due to increases in particulate matter. Ethanol from sugar cane, however, is much more efficient than ethanol from corn, and ethanol from cellulose is another promising and rapidly developing alternative (see Hill et al., 2009, for more on the differences among various sources of and processes for producing ethanol).

However they respond to a gasoline tax, consumers will choose the cheapest of the set of all behaviors that reduce miles driven or increase miles per gallon. The same would be true if they were faced with a carbon tax or if producers were required to pay a fee for a carbon permit. Direct taxes on miles driven or PAYD insurance would also reduce miles driven and lead to reductions in gasoline consumption. While these mileage fees could efficiently internalize the per mile externalities discussed above, they have been challenged by some who suggest that consumers may tamper with their odometers or that calculating miles driven using GPS technology may infringe on civil liberties.

On the other hand, any policy that acts only to increase fuel efficiency involves the undesirable incentive to drive more miles. Increases in fuel economy reduce the cost per mile, because they increase the number of miles that can be driven per gallon. As with any good, the law of demand holds here—as the price of miles drops, the quantity of miles driven rises.

The United States' main policy to increase fuel economy is Corporate Average Fuel Economy (CAFE) standards. It also assesses so-called gas guzzler taxes, but only on the most inefficient passenger cars, which are exclusively luxury sports cars. CAFE standards mandate that each vehicle manufacturer attain a specific average fuel economy in the cars it sells, and another applies for light trucks. At the time this book went to press, the federal government was in the process of overhauling these rules. The new rules will almost certainly impose more stringent (higher) fuel economy requirements.

CAFE is a CAC regulation. In particular, CAFE is a performance-based standard that allows manufacturers to choose the method by which to meet the standard. As such, performance-based standards are more flexible than a technology standard, which mandates that all manufacturers supply specific technologies (such as electric vehicles).

Still, once automakers have reached the level specified by the standard, they face little incentive to increase fuel economy any further. And CAFE applies a one-size-fits-all policy to firms that may differ widely in size and cost structure. Thus, while fuel economy is increased to the level desired by policy makers (which may be inefficiently low or high), it is accomplished at a higher cost to firms and consumers than might have occurred if different firms were allowed to attain different levels of average fuel economy. If firms for whom innovation in fuel economy technology is easier were induced to sell more vehicles than firms for whom such technology is more difficult, fuel economy goals could be attained at lower overall costs to society.

Feebates, which subsidize the purchase or manufacture of fuel-efficient vehicles and tax fuel-inefficient vehicles, would solve this problem, because manufacturers could choose exactly how to respond to the policy. Those who innovate cheaply would respond more than those for whom innovation is costly. A tax on fuel-inefficient vehicles would also be efficient, but the feebate has the potential advantage of being inherently revenue neutral, so any tax revenues taken in by the program would be rebated in the

form of subsidies to fuel efficiency. Such revenue neutrality is generally more attractive to policy makers in the United States, who risk damage to their popularity when they increase the overall tax burden placed on their constituents.

Both CAFE and feebates, however, increase miles driven by reducing the price per mile. This is the rebound effect, where gasoline consumption rebounds because driving is cheaper, and offsets some of the reduction in gasoline made possible by higher fuel economy. Ken Small and Kurt Van Dender (2007) estimate this effect is about 10% of the initial reduction in gasoline consumption due to more stringent CAFE standards. Unless consumers fail to buy or producers fail to offer the efficient amount of fuel economy for reasons other than the existence of the external costs discussed above, then gasoline taxes, carbon taxes, or carbon cap-and-trade programs are almost certainly more efficient than policies that increase fuel economy but also increase miles driven.

## Failures in the Market for Fuel Economy

The conclusion that a gasoline tax is an efficient (and thus preferable) policy instrument for the reduction of gasoline consumption rests on an economic model in which consumers make fully informed, rational decisions regarding both how many miles to drive and what vehicle to buy. It also requires that automakers provide consumers with an efficiently varied set of vehicles from which to choose. There are reasons to think that these conditions may not hold. In particular, there is evidence that consumers may not sufficiently value fuel economy when making the decision about what car to buy, and logical arguments suggest that producers offer fewer than the optimal number of choices of fuel efficiencies.

### Possibilities for Market Failure Due to Consumer Fallibility

Consumers may undervalue fuel economy because they do not fully and correctly calculate the present discounted value of the flow of future operating costs, which depend on driving behavior, fuel economy, and the future price of gasoline. That is, when choosing between a more fuel-efficient and a less fuel-efficient vehicle, a consumer may not correctly calculate or perceive the difference in what he or she will pay for gasoline in the one vehicle versus the other over the entire period during which he or she will own the vehicle.

This implies that in the absence of a corrective policy, consumers will buy too little fuel economy. It also implies that consumers will underreact to a greenhouse gas policy that raises the value of fuel economy by increasing the price of gasoline. As such, an additional policy, such as the CAFE standard, or a feebate, which subsidizes the purchase of fuel-efficient vehicles and taxes the purchase of inefficient vehicles, can be welfare improving, even when carbon emissions have been priced optimally via a gasoline tax, a carbon tax, or a cap-and-trade program (see Fischer, Harrington, & Parry, 2007; Greene, Patterson, Singh, & Li, 2005; National Research Council, 2002; Train, Davis, & Levine, 1997, for discussions of the possibility that CAFE or feebates may be efficient in the presence of such consumer fallibility).

A growing body of evidence suggests many reasons to think that consumers may make systematic mistakes regarding the value of fuel economy. First, experimental evidence suggests that consumers are confused by the nonlinearity of cost savings in miles-per-gallon-rated fuel economy (Larrick & Soll, 2008). For example, an increase in fuel economy from 15 to 16 miles per gallon implies a greater cost savings than an increase from 30 to 31 miles per gallon, but consumers tend to think these changes result in the same cost savings. Second, qualitative evidence from surveys of new car buyers suggests that most consumers are unable to articulate the key building blocks required for a present discounted-value analysis, including typical mileage, fuel economy ratings of their current vehicles, and a discount rate (Turrentine & Kurani, 2007). Third, analysis of data from the Michigan Survey of Consumers produces several results that are difficult to explain in a model of infallible consumers. Each month, the survey asks a sample of households whether it is a good time to buy a vehicle and why. Even when controlling for the real price of gasoline (the price that consumers pay, adjusted by inflation to account for the changes in purchasing power over time), consumers are more likely to cite gasoline prices and fuel economy as important when nominal gasoline prices (the price that is posted on gasoline station signs) are high and when prices have just changed. They also appear to respond more strongly to gasoline price increases than decreases (Sallee & West, 2007). There is also evidence that consumers act the same way when buying other durable goods, such as refrigerators and air conditioners—they do not fully account for the differences in energy costs when deciding between buying a more expensive but more fuel-efficient appliance and a cheaper but less fuel-efficient appliance (Dubin & McFadden, 1984; Hausman, 1979).

Skeptics, however, point out that of a vehicle's many characteristics, consumers probably understand fuel economy better than most of its other characteristics (Kleit, 2004). After all, a vehicle's fuel economy rating appears in large type on a window sticker placed on every new vehicle sold in the United States. And other researchers have found that consumers buying used cars do seem to fully account for differences in future operating costs across vehicle models (Dreyfus & Viscusi, 1995).

Part of the reason that economists have yet to reach an agreement on the existence or magnitude of consumer myopia or fallibility is that until recently, data and computing requirements were too onerous. To properly identify consumer fallibility, one must isolate consumer

responses from a whole host of other factors, including broad economy-wide movements and a multitude of production and pricing behaviors on the part of automakers. Work that measures the effects of changes in gasoline prices in used-car markets is promising, because it can more easily control for the confounding effects of vehicle manufacturers' pricing decisions. Such work finds that consumers internalize a surprisingly small proportion of operating costs when deciding which used car to buy (see Kahn, 1986; Kilian & Sims, 2006; Sallee & West, 2009) and provides preliminary support for the notion that consumers are fallible.

### Possibilities for Market Failure Due to Producer Behavior

It is also possible that market failures exist on the producer side. But even less is known about producer behavior, especially because (quite understandably) automakers are hesitant to share information about their production and pricing strategies.

Portney, Parry, Gruenspecht, and Harrington (2003) list a number of reasons why producers may undersupply fuel economy. First, the vehicle industry is oligopolistic, in that a relatively small number of firms dominate the market. This may imply that these firms do not face the same incentives to provide fuel economy as firms in more competitive markets. Instead, these companies may engage in strategic behaviors that result in a lower-than-optimal amount of innovation.

Second, although engineering studies indicate that it is in the interest of automakers to supply vehicles of greater fuel economy, such studies may not capture the full costs of developing and implementing a new technology. Third, if automakers are unable to reap the full returns of investment in fuel-saving technologies because of copycat activity on the part of competitors, they will invest in less technology than is socially optimal.

However, it may simply be the case that consumers are unwilling to forgo the performance attributes, such as acceleration, that would likely be lost if automakers were to sell models that are more fuel efficient. Automakers may simply be responding optimally to consumer demands. After all, as Portney et al. (2003) remind readers, there is already a wide variety of fuel-efficient models from which consumers can choose.

### Policy Implications if the Market for Fuel Economy Is Inefficient

Findings of consumer fallibility or inefficient producer behavior have the potential to shake the world of environmental policy making, because they imply that a policy that sets a price on carbon will not be sufficient to efficiently reduce gasoline consumption, and therefore greenhouse gases, from the personal transportation sector.

If consumers or producers do not buy or supply an efficiently high amount of fuel economy, then to attain the efficient amount of gasoline consumption, a tax or fee on carbon or a tax on gasoline must be paired with another incentive that induces further improvements in fuel economy. If consumers are fallible or producers provide too little fuel economy because of strategic behavior, the increases in miles driven that come about from increases in fuel economy will be warranted.

For the reasons discussed above, a feebate—or even just a tax on fuel-inefficient vehicles—would be more efficient than making CAFE standards more stringent. If consumers are myopic enough, disregarding part or all of the future fuel cost savings from choosing a more fuel-efficient vehicle, making CAFE more stringent might increase efficiency. The same is true if producers' supply choices are sufficiently constrained by strategic or other considerations.

## Distributional Concerns

Because the extent and the magnitude of externalities in gasoline markets are so clear, most economists focus their attention on the potential efficiency-improving characteristics of policies that internalize such externalities. Equity, or distributional, implications of a tax on gasoline or carbon, however, are often cited as one of the strongest arguments against increasing or imposing such taxes. Many worry that poor households will bear a greater burden because expenditure on gasoline tax would be a larger proportion of their incomes than it would be for wealthy households. That is, they worry that a gasoline tax is regressive, as opposed to progressive, where the amount paid in tax as a proportion of income increases as income increases. They also worry that households will bear more of the burden than producers.

As explained in West and Williams (2004), an ideal measure of tax incidence would begin with a calculation of all of the changes in prices that would occur throughout the economy in response to the change in the tax rate and proceed with a calculation of the effects of those price changes on households' well-being. The obvious effect of an increase in the gasoline tax is a rise in the consumer price of gasoline, imposing a burden on gasoline consumers. But an increase in the gasoline tax could also lower the producer price of gasoline, imposing a burden on the owners of gas stations, gasoline producers, and perhaps in turn, through lower wages, on workers in those industries. It could also affect the prices of other goods that use gasoline as an intermediate input.

However, calculating such effects requires a great deal of information, most notably the demand and supply elasticities for all relevant industries and the distribution of ownership of firms in those industries. A few papers either hazard a guess on the relative burden on producers and consumers (e.g., Congressional Budget Office, 2003) or

estimate the relative burdens explicitly (e.g., Skidmore, Peltier, & Alm, 2005). But most incidence studies assume that the supply of consumer goods is perfectly elastic. This implies that the imposition of a tax on a consumer good does not affect the producer price of that good and thus that the entire burden of the tax falls on consumers.

These studies find that the gasoline tax is indeed quite regressive, one of the most regressive taxes in the economy (see Poterba, 1991; West, 2004; West & Williams, 2004). This is not surprising, because gasoline, like all energy, is a necessity—as income increases, so does consumption of gasoline, but not by enough to make wealthy households spend more on it as a proportion of their income than do poor people.

But West and Williams (2004) suggest a simple way to mitigate or even completely overcome the regressivity of the gasoline tax. By using the revenues from the gasoline tax to reduce taxes on work, the policy can be made significantly less regressive. And by using the revenues to give rebates of the same amount to all households, the policy can actually be made progressive. Such rebates do not reduce the efficiency of the gasoline tax, because the tax still provides the incentive to reduce gasoline consumption. Indeed, if gasoline tax revenues are used to reduce labor taxes, the overall efficiency of the gas tax policy is improved.

In part because such regressivity-reducing revenue rebate schemes are available, but more importantly because the externalities from gasoline use are substantial, economists do not rule out taxes on gasoline because of equity concerns. Indeed, more often, they operate on a general principle that it is best to use environmental policies to internalize externalities, regardless of their distributional implications, while relying on the income tax system to attain equity goals.

Unfortunately, very little is known about the equity implications of CAFE standards. When CAFE standards are binding, making automakers do something they would not have otherwise done, they increase the price of new vehicles because they force automakers to incur costs when investing in fuel-saving vehicle technologies. Because wealthy households are more likely than poor households to buy new vehicles, one might posit that wealthy households would pay the higher prices for new vehicles. In this case, CAFE standards would be progressive. But autoworkers are likely to bear some of the burden of binding CAFE standards, and used-car markets are also likely to be affected by CAFE, making the overall incidence of CAFE uncertain.

Feebates, on the other hand, would make new fuel-efficient vehicles cheaper. Both consumers and producers might be expected to reap some of the benefits of such a subsidy, and those selling and those buying gas guzzlers would share the costs of the tax portion of the policy. To the extent that wealthy households buy more fuel-efficient vehicles than poor households and therefore are more likely to receive a government subsidy, a feebate might be expected to be regressive.

## Conclusion and Future Directions

If consumers are making fully informed, rational decisions about what kind of vehicles to drive and producers both reap the full benefits of investment in fuel-saving technologies and act competitively when supplying fuel economy, then policy makers can rely on a gasoline tax, a carbon tax, or a carbon cap-and-trade program to efficiently reduce greenhouse gases from the personal transportation sector. In such a context, it is very unlikely that CAFE standards, now the primary policy used to encourage fuel saving in the United States, can be efficient alone or in conjunction with a tax on gasoline or carbon policy. The primary problem with CAFE is that it reduces the price of driving, thereby inducing consumers to impose greater congestion and accident costs on others. Because a feebate also reduces the price of driving, it too can be expected to be inefficient.

Only if some additional imperfection causes the market for fuel economy to fail, if consumers are short sighted, or if automakers are restricted or restrict themselves in supplying the optimal variety of fuel efficiencies, can we expect for CAFE or a feebate to improve efficiency. If the market for fuel economy fails, economic theory tells us that the one-size-fits-all nature of CAFE, at least as the policy currently stands, renders it less preferable than a feebate.

To determine whether a tax on gasoline, a carbon tax, or a carbon cap-and-trade program is sufficient to induce the optimal amount of gasoline consumption and greenhouse gas emissions from the transportation sector—and how to best implement such policies—researchers must now focus on three main areas. First, they need to identify and quantify potential failures in the market for fuel economy. Second, they should refine estimates of the external costs of gasoline as new information on the potential effects of climate change becomes available. Third, they should gain a greater understanding of the potential distributional effects of any policy designed to increase fuel economy.

## References and Further Readings

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics, 3*(1), 1–44.

Congressional Budget Office. (2003). *The economic costs of fuel economy standards versus a gasoline tax.* Washington, DC: Author.

Davis, S. (2004). *Transportation energy data book* (24th ed.). Oak Ridge, TN: Oak Ridge National Laboratory.

Davis, S. (2007). *Transportation energy data book* (27th ed.). Oak Ridge, TN: Oak Ridge National Laboratory.

Dreyfus, M. K., & Viscusi, W. K. (1995). Rates of time preference and consumer valuations of automobile safety and fuel efficiency. *Journal of Law and Economics, 38,* 79–105.

Dubin, J. A., & McFadden, D. L. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica, 52*(2), 345–362.

Federal Highway Administration. (2009). *Traffic volume trends.* Available at http://www.fhwa.dot.gov/ohim/tvtw/tvtpage.cfm

Fischer, C., Harrington, W., & Parry, I. (2007). Do market failures justify tightening corporate average fuel economy (CAFE) standards? *Energy Journal, 28*(4), 1–30.

Greene, D. L., Patterson, P. D., Singh, M., & Li, J. (2005). Feebates, rebates and gas-guzzler taxes: A study of incentives for increased fuel economy. *Energy Policy, 33*(6), 757–775.

Hausman, J. A. (1979). Individual discount rates and the purchase and utilization of energy-using durables. *Bell Journal of Economics, 10*(1), 33–54.

Hill, J., Polasky, S., Nelson, E., Tilman, D., Huo, H., Ludwig, L., et al. (2009). Climate change and health costs of air emissions from biofuels and gasoline. *Proceedings of the National Academy of Sciences, USA, 106,* 2077–2082.

Jaffe, A., & Stavins, R. (1995). Dynamic incentives of environmental regulations: The effects of alternative policy instruments on technology diffusion. *Journal of Environmental Economics and Management, 29,* S43–S63.

Kahn, J. A. (1986). Gasoline prices and the used automobile market: A rational expectations asset price approach. *Quarterly Journal of Economics, 101*(2), 323–340.

Kilian, L., & Sims, E. R. (2006). *The effects of real gasoline prices on automobile demand: A structural analysis using micro data.* Unpublished manuscript, University of Michigan.

Kleit, A. N. (2004). Impacts of long-range increases in the corporate average fuel economy (CAFE) standard. *Economic Inquiry, 42*(2), 279–294.

Laffont, J., & Tirole, J. (1996). Pollution permits and environmental innovation. *Journal of Public Economics, 62,* 127–140.

Larrick, R. P., & Soll, J. B. (2008). The MPG illusion. *Science, 320*(5883), 1593–1594.

Li, S., Timmins, C., & von Haefen, R. H. (2009). How do gasoline prices affect fleet fuel economy? *American Economic Journal: Economic Policy, 1*(2), 113–137.

National Research Council. (2002). *Effectiveness and impact of corporate average fuel economy (CAFE) standards.* Washington, DC: National Academy of Sciences.

Parry, I. W. (1998). Pollution regulation and the efficiency gains from technology innovation. *Journal of Regulatory Economics, 14,* 229–254.

Parry, I. W. (2004). Comparing alternative policies to reduce traffic accidents. *Journal of Urban Economics, 54*(2), 346–368.

Parry, I. W. (2005, April). *Is pay-as-you-drive insurance a better way to reduce gasoline than gasoline taxes?* (Resources for the Future Discussion Paper 0515). Washington, DC: Resources for the Future.

Parry, I. W., Walls, M., & Harrington, W. (2007). Automobile externalities and policies. *Journal of Economic Literature, 45,* 374–400.

Portney, P., Parry, I. W., Gruenspecht, H. K., & Harrington, W. (2003). The economics of fuel economy standards. *Journal of Economic Perspectives, 17*(4), 203–217.

Poterba, J. M. (1991). Is the gasoline tax regressive? In D. Bradford (Ed.), *Tax policy and the economy 5* (pp. 145–164). Boston: MIT Press.

Sallee, J. M., & West, S. E. (2007). *The effect of gasoline price changes on the demand for fuel economy.* Unpublished manuscript.

Sallee, J. M., & West, S. E. (2009, January). *Testing for consumer myopia: The effect of gasoline price changes on the demand for fuel economy in used vehicles.* Paper presented at the meeting of the Allied Social Science Associations, San Francisco.

Small, K., & Van Dender, K. (2007). Fuel efficiency and motor vehicle travel: The declining rebound effect. *Energy Journal, 28*(1), 25–51.

Skidmore, M., Peltier, J., & Alm, J. (2005). Do motor fuel sales-below-cost laws lower prices? *Journal of Urban Economics, 57*(1), 189–211.

Stavins, R. N. (2000). Market-based environmental policies. In P. Portney & R. N. Stavins (Eds.), *Public policies for environmental protection* (pp. 31–76). Washington, DC: Resources for the Future.

Train, K. E., Davis, W. B., & Levine, M. D. (1997). Fees and rebates on new vehicles: Impacts on fuel efficiency, carbon dioxide emissions, and consumer surplus. *Transportation Research Part E: Logistics and Transportation Review, 33*(1), 1–13.

Turrentine, T. S., & Kurani, K. S. (2007). Car buyers and fuel economy. *Energy Policy, 35*(2), 1213–1223.

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. (2004). Estimates from a consumer demand system: Implications for the incidence of environmental taxes. *Journal of Environmental Economics and Management, 47,* 535–558.

West, S. E., & Williams, R. C., III. (2005). The cost of reducing gasoline consumption. *American Economic Review, 95,* 294–299.

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.

West, S. E., & Wolverton, A. (2005). Market-based policies for pollution control in Latin America. In A. Romero & S. West (Eds.), *Environmental issues in Latin America and the Caribbean* (pp. 121–148). Dordrecht, the Netherlands: Springer Press.

# 25

# Public Finance

Neil Canaday
*Wesleyan University*

Public finance, commonly referred to as public economics, is the field of economics that examines the role of the government in the economy and the economic consequences of the government's actions.[1] A notable exception to this definition is the study of the government's effects on the business cycle, which is usually considered to be a part of macroeconomics rather than public finance. Public finance is concerned with both positive and normative economic issues. Positive economics is the division of economics that examines the consequences of economic actions and thus includes the development of theory, whereas normative economics brings in value judgments about what should be done to analyses and often gives recommendations for public policy. Normative economic issues, in fact, are discussed and debated more often in the public finance literature than in the literatures of most other fields in economics.[2]

Public finance covers a wide range of topics, many of which are central to the economics discipline. The public finance literature examines theoretical as well as empirical topics. Some studies focus on macro-issues, and others examine specific economic issues. Notwithstanding this diversity in coverage, most questions addressed by the public finance literature fall into one or more of the following three categories: (1) Under what scenarios should governments intervene in the economy, (2) what are the economic consequences of government interventions in the economy, and (3) why do governments do what they do?

Our present knowledge of public finance is the result of more than 200 years of scholarly contributions. Although originally published in 1776, *The Wealth of Nations* by Adam Smith popularized many notions about the proper role of the government that are still very much relevant to today's world. Adam Smith discussed the following three duties that the government should perform: (1) protect its citizens from foreign invaders by providing national defense; (2) protect members of society from injustices from each other through the provision of a legal system; and (3) provide certain public works, such as bridges, roads, and institutions, including elementary level education (but not secondary and university education).[3] Most people, including economists, would agree with Adam Smith that the government should perform these three tasks. Of course, many people believe the government should do even more. Many readers of this chapter may strongly believe the government should provide for the education of students at universities and colleges.

As the role of the government in the economy has changed over time, the focus of the public finance literature has similarly evolved. In the 1950s and 1960s, the emphasis of public finance was largely on issues of taxation. Now, with the government significantly involved in many aspects of the economy, the public finance literature has expanded its focus to include virtually all facets of government spending, as well as taxation. Many advances have been made within the field of public finance over the past several decades, and public finance economists have made substantial contributions to many other fields in economics. For example, the economics of aging, a relatively new economic subfield, has benefited greatly from public finance economists who have provided analysis and policy recommendations for issues pertaining to government entitlement programs for retirees. Many of these topics are of particular importance to Americans. How will the approaching retirement of the baby boomer generation affect the overall level of economic activity? And how will this growth in retirees

strain government entitlement programs such as Social Security and Medicare, and what should the government do, if anything, to these entitlement programs as a result?

## Public Finance Theory

### The Theory of the Government

When the proper role of the government is under consideration, it is often useful to think about under what circumstances the government can do a better job than private markets. If reliance on private markets' carrying of economic activities always results in economically efficient outcomes and people are, for the most part, content with the distribution of income and wealth, then there is little or no need for a government role in the economy. However, there are many circumstances where the private market fails to achieve an efficient outcome or the private market outcome results in a distribution of income or wealth that is considered by most people to be very unjust. Sometimes there are market failures. In these situations, society may be better off when actions are conducted by the government, rather than letting economic activities take place solely in private markets.

One task the government can likely do better than private markets is the provision of so-called public goods. The theory of public goods was to a large extent developed by the economist Paul Samuelson.[4] To economists, public goods are not the same things as goods provided by the public sector. Rather, public goods are goods that have the following two qualities: (1) nonrivalry and (2) nonexcludability. In comparison, private goods have both rivalry and excludability. Nonrivalrous goods are goods for which consumption by one consumer does not prevent or diminish simultaneous consumption by other consumers. On the other hand, a good has rivalry if consumption by one consumer prevents or diminishes the ability of other consumers to consume it at the same time. A slice of pizza is clearly a rivalrous good. Nonexcludability means it is not possible to prevent people who have not paid for it from enjoying the benefits of it. Air is an example of a nonexcludable good. National defense is a public good, because it has both nonrivalry and nonexcludability in consumption. Although the government provides many public goods, it also provides many private goods, such as health care, school lunches, and Social Security.

The free-rider problem is a primary cause of private markets' failure to efficiently provide public goods. Because people benefit from public goods regardless of who in society pays the costs of providing them, people have an incentive to spend too little on public goods themselves and instead take a free ride on the contributions of others. Therefore, when private markets provide public goods, too little is often spent on their provision. The government can correct this market inefficiency by taxing members of society and using this tax revenue to provide an efficient amount of public goods. On the other hand, taxes also cause market inefficiencies, a matter that is discussed later in this chapter.

The problems associated with the provision of public goods can alternatively be discussed in the framework of private and social costs and benefits. The marginal private benefit is the incremental benefit of an activity for private individuals or businesses engaged in that activity, whereas the marginal social benefit measures the incremental benefit of an activity for society. Marginal private costs and marginal social costs are similarly defined. Therefore, the demand and supply curves are the same as marginal private benefit and marginal private cost curves, respectively. Public goods create a positive externality, which is to say that public goods provide external benefits. The external benefit or positive externality is measured by the amount to which social benefits exceed private benefits. In other situations, there are negative externalities. A negative externality exists when the social cost exceeds the private cost.

Figure 25.1 illustrates the case for a positive externality. When the provision of the public good is left to private markets, individuals will purchase units of the public good until their marginal private costs are equal to their marginal private benefits. In other words, marginal private cost is the supply curve, and marginal private benefit is the demand curve. For the overall market, the quantity of the public good provided, $Q^c$, at the competitive equilibrium is not economically efficient. More specifically, too little of the public good is provided by the private market. Notice that at the competitive equilibrium (where supply equals demand), the marginal social benefit still exceeds the marginal social cost. The economically efficient outcome, $Q^*$, occurs where the marginal social benefit is equal to the marginal social cost. There are potential social gains to government intervention if the government's actions result in an increase in the quantity of the public good that is closer to the economically efficient outcome than the competitive equilibrium. For example, $Q^G$ in Figure 25.1 is one possible outcome of government intervention that results in a more economically efficient outcome than the competitive equilibrium.

Under certain circumstances, government intervention is not needed to correct the externality problem because the private market will be able to solve the externality problem on its own and provide an economically efficient outcome. The Coase Theorem, attributed to Ronald Coase, provides the conditions where the market may work efficiently even if externalities are present. The Coase Theorem states that when property rights are well defined and the transaction costs involved in the bargaining process between the parties are sufficiently low, the private market may provide an economically efficient outcome even though externalities are present.

How the Coase Theorem works can be illustrated with a simple example. Suppose Jack and Jill live next door to each other and Jack is contemplating planting a flower garden in his front yard. The flower garden will cost Jack $50 to plant (because of seeds, water, fertilizer, and labor) and

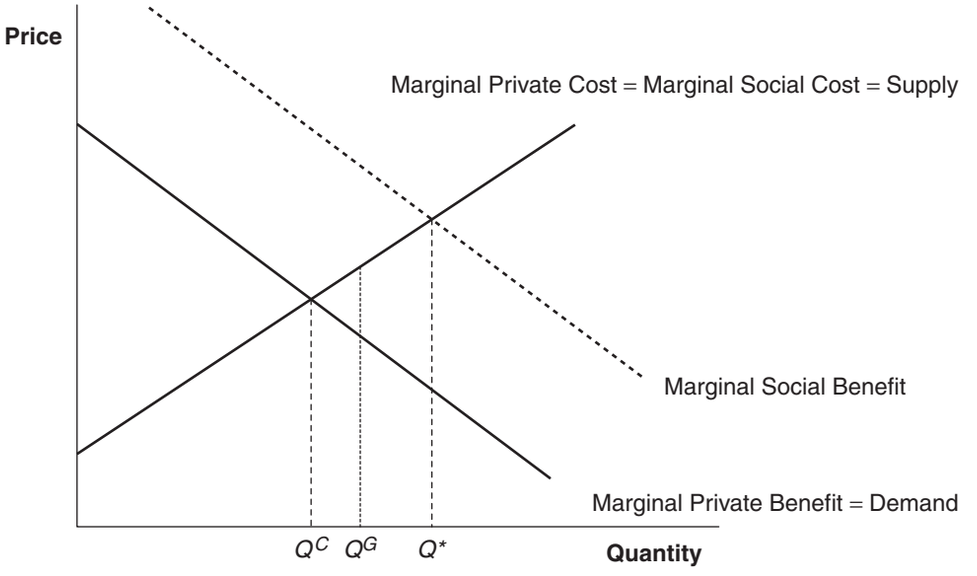


**Figure 25.1** Effects of a Tax

will provide him a benefit equal to $40 because of its beauty. The flower garden will also provide an external benefit to Jill equal to $20. This positive externality arises because Jill will get to enjoy the beauty of the flower garden even though she does not own it. Clearly, from a social viewpoint, the flower garden should be planted, because it has a total social benefit of $60, which exceeds its total social cost of $40. Nonetheless, Jack will not plant the flower garden if he has to rely solely on his own funding, because his private benefit of $40 is less than his private cost of $50. The first condition of the Coase Theorem—that property rights are well defined—is met, because Jack has the property rights over whether to plant the flower garden. Suppose further that Jack and Jill can bargain over the planting of the flower garden at zero cost. With this assumption, the second condition of the Coase Theorem—that the transaction costs of bargaining are sufficiently low—is also met.

Now, let us see the Coase Theorem in action: Jill would be willing to pay Jack an amount that is sufficiently high to induce him to plant the flower garden. As a result of this side payment, both Jack and Jill would be better off, because the flower garden will be planted and the efficient economic outcome will arise. For example, if Jill pays Jack $15 to plant the flower garden, her total benefit of $20 exceeds her total cost of $15. Similarly, for Jack, his total benefit would be $55 ($40 from getting to enjoy the flower garden and $15 from Jill), which exceeds his total cost of $50.

When only a few parties are involved in situations of externalities, the transaction costs of bargaining may very well be sufficiently low, like in the above example, and private markets may be able to correct the problem of externalities on their own, providing an economically efficient outcome. In fact, judges and legal scholars sometimes use the Coase Theorem as a guide for thinking about proper solutions to tort cases involving public nuisances.

Unfortunately, many serious situations of externalities involve many parties, and the costs of bargaining are prohibitively high. For example, situations where companies are polluting the air or water often involve many victims, and bargaining between all the parties involved would be a very costly and complicated process. It is farfetched to think that private markets can solve problems of externalities and achieve economically efficient outcomes on their own in situations of industrial pollution. In these cases, government intervention in the marketplace could improve on the private market outcome. In fact, government intervention in situations where there is degradation of the environment is common.

So far, the discussion of the government has focused on what the proper role of the government should be and how the government can improve on outcomes left solely to actions of private markets. However, economists are also interested in why governments do what they do. The subdiscipline of public finance called public choice tries to explain how democratic governments actually work, rather than how they should work. *The Calculus of Consent: Logical Foundations of Constitutional Democracy,* coauthored by James M. Buchanan and Gordon Tullock (1962), is considered by many economists the landmark text that founded public choice as a field in economics. Although public choice shares many similarities with political science, public choice economists assume that voters, politicians, and bureaucrats all act predominately in their own self-interest, rather than in the interests of the public good.

## The Theory of Taxation

A natural starting point for examining the modern theory of taxation is Richard Musgrave's 1959 textbook, *The Theory of Public Finance.* Musgrave's public finance text

was the first to look like a present-day economics textbook, using many algebraic equations and graphs to illustrate the economic effects of government policy. Musgrave's textbook provides a thorough partial equilibrium analysis of taxation. Partial equilibrium analysis examines the equilibrium in one market without factoring in ripple effects on other markets. In comparison, general equilibrium analysis examines the entire economy, and therefore, it takes into account cross-market effects. Two important economic questions of taxation are (1) what is the incidence of the tax and (2) how large is the loss in economic efficiency from the tax? The deadweight loss of the tax is another term for the economic efficiency loss of the tax. Tax incidence measures how the burden of a tax falls on different members of society. In practically all cases, taxes create economic inefficiencies through the distortion of incentives.

Figure 25.2 illustrates the concepts of tax incidence and deadweight loss within the partial equilibrium framework. Without a tax, the competitive equilibrium occurs where the demand equals supply at a price of $P^*$ and a quantity of $Q^*$. The equilibrium with no tax is economically efficient because the marginal social cost is equal to the marginal social benefit. Now, suppose a tax is imposed. The tax shifts up the supply curve vertically by the amount of the tax per unit.[5] The equilibrium quantity with the tax, $Q^T$, is below the equilibrium quantity without the tax, $Q^*$. The tax distorts market behavior in an inefficient manner, which is evident because the units between $Q^T$ and $Q^*$ are not being bought and sold when the tax is imposed, even though the marginal social benefits of these units exceed their marginal social costs. The deadweight loss of tax is represented by the triangle shaded in gray. Notice that the tax puts a wedge between the price consumers pay, $P^C$, and the price sellers get to keep net of the tax, $P^S$. The tax revenue per unit that the government receives is measured by the difference between the consumer price and seller price $(P^C - P^S)$.

The tax incidence is examined by comparing how much of the tax falls on consumers and how much falls on sellers. The dollar amount of the tax per unit falling on consumers is equal to the price consumers pay, with the tax subtracted from the price they pay without the tax $(P^C - P^*)$. Similarly, the dollar amount of the tax per unit falling on sellers is equal to the price sellers receive, without the tax subtracted from the price sellers receive with the tax $(P^* - P^S)$. The addition of the dollar amount of the tax per unit falling on consumers to the dollar amount falling on sellers equals the amount of the tax per unit $(P^C - P^S)$. In the example illustrated in Figure 25.2, the tax burden falls disproportionately on consumers.

It is important to note that how much of the tax collected from consumers, as opposed to sellers, has absolutely no effect on either the incidence or the deadweight loss of the tax. In Figure 25.2, the tax is levied entirely on sellers, as is evident by the tax raising the marginal cost exactly by the amount of tax per unit. Alternatively, if the tax were levied entirely on consumers, the demand curve would have shifted vertically downward by the amount of the tax per unit. In both cases, the consumers' net price, the sellers' net price, and the size of the deadweight loss are all exactly the same.

The elasticity of demand and the elasticity of supply are the factors that determine the incidence and deadweight loss of a tax. Some generalizations can be made about how the elasticity of demand and supply influence the tax incidence and the size of the deadweight loss of a tax. The more inelastic the demand or the supply, all else equal, the larger the share of the tax burden that falls on consumers and the smaller the share that falls on sellers.

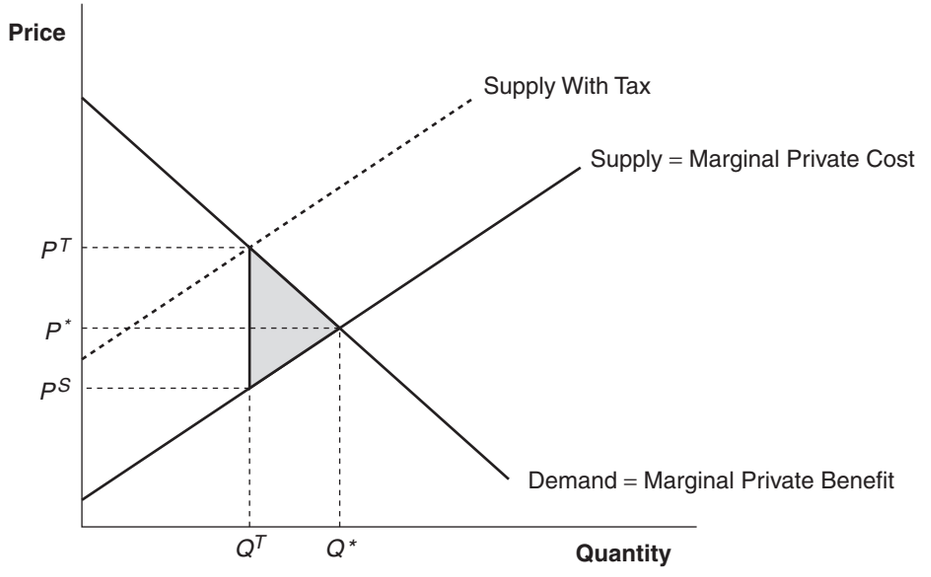


**Figure 25.2** External Private Benefits

Additionally, the more elastic the demand or the supply, the larger the deadweight loss of the tax will be. For example, suppose a tax is imposed on gasoline, a product with a relatively inelastic demand. The burden of this tax will fall primarily on consumers of gasoline, even though governments collect this tax from sellers. Furthermore, because the demand for gasoline is relatively inelastic, the deadweight loss of the tax will be relatively small because there will be a relatively small change in the quantity of gasoline bought and sold.[6] Although partial equilibrium analysis is helpful for understanding many issues of taxation, it is inadequate for providing profound insight about some important economic issues surrounding taxation. For instance, partial equilibrium analysis often cannot adequately describe the incidence and deadweight loss of taxes for situations where there are multiple products and multiple sectors of the economy. Additionally, partial equilibrium analysis is not useful, in most cases, for determining the optimal tax structure of the tax system (e.g., how much of the tax rates should be on wages, savings, and different products). To examine these issues of taxation, public finance economists rely on general equilibrium models.[7]

The scholarly work by Arnold Harberger shows how the techniques used to measure the incidence and deadweight loss of taxes in the partial equilibrium framework can be extended to situations with multiple products and sectors, including the case of the corporate income tax. The deadweight loss resulting from a tax or some other type of distortion, such as a price floor or a price ceiling, is often referred to as Harberger's Triangle. In Figure 25.2, Harberger's Triangle is the triangle shaded in gray.

The incidence of a tax may differ widely depending on whether the economy is opened or closed (i.e., whether the economy, for the most part, freely trades with other economies).[8] Harberger (1962) examines the corporate tax incidence with a general equilibrium model of a closed economy that has two sectors (corporate and noncorporate) and two factors (labor and capital). Harberger (2008) has recently revisited his work on the incidence of the corporate tax. For the interesting case where product demands and the production functions in both sectors conform to the commonly used Cobb-Douglas functional form, the entire burden of the corporate income tax falls on capital. In comparison, models that assume the economy is opened often find that much of the burden of the corporate income tax is shifted to labor.

Economic theories have addressed the design of an optimal system of taxation, a topic related to the measurement of the deadweight loss of taxes. Although taxes create losses in economic efficiency by distorting people's behavior, they are needed to generate revenues for a society to pursue its social objectives. Theories of optimal taxation look for the system of taxes that minimizes the economic efficiency costs of taxes, subject to constraints, such as the types of taxes and information that are available to the government. Theoretical inquiries into optimal taxation fall into one of the following three categories: (1) the design of the optimal system of commodity taxes; (2) the design of a general system of taxation, including nonlinear taxes on income, with a focus on the role of taxes for addressing concerns about inequality; and (3) the role of taxes to address market failures. The first strand began with the work by Ramsey (1927) and has been added to, most notably, by Peter Diamond and James Mirrlees (1971a, 1971b). The Ramsey tax rule says that under a tax system composed only of commodity taxes, the optimal tax rates are higher for commodities with more inelastic demands. Thus, the optimal tax rate is higher for gasoline than for pretzels, assuming that gasoline is more inelastically demanded than pretzels. The work by Mirrlees (1971) was the first in the second strand of theory. Finally, the seminal work by Arthur Pigou (1947) provides the foundation for work in the third strand of theory. A Pigouvian tax is levied to correct a negative externality. For example, suppose that the production of widgets has a constant $1 external cost. The government can fully correct this negative externality by imposing a constant $1 tax per widget. When this tax is imposed, widget producers will act as if their private costs are equal to the social costs, and the economically efficient outcome will prevail.

The extent to which economic inputs into production, such as labor and capital, are mobile has significant implications for taxation as well as for many other aspects of public finance. Economists have developed theoretical models to examine the effects of tax competition—a situation where governments lower taxes or provide other benefits in an effort to encourage productive resources to relocate within their borders—on tax rates, tax incidence, and on the distribution of resources across regions. Tax competition can occur between countries, states, or smaller units of government.

Also related to factor mobility is the Tiebout model, developed by Charles Tiebout (1956). Tiebout's insight is that the free-rider problem pertaining to public goods is different in the context of local governments because they offer bundles of goods and services and taxes to potential residents because people are able to take into account, when deciding where they want to live, the governmental amenities and disamenities offered by the various localities. In the Tiebout model, competition among local governments and people's ability to vote with their feet effectively solves the free-rider problem and results in an economically efficient provision of public goods by local governments. When people have different tastes for amenities, the Tiebout model predicts that there will be locational sorting based on individuals' preferences. For example, people with children will choose to live in locations with high tax rates for schools and well-funded public schools, whereas people without children, such as retirees, will choose to live in locations with low school tax rates and poorly funded public schools.

## Applications and Empirical Research

The empirical public finance literature has made significant strides over the past several decades. Several factors, including advances in computer technology, econometric techniques, and software, have contributed to improvements in both the quality of research and to the exploration of new avenues of research in public finance. Additionally, new data sets containing valuable information at the individual and household levels have been created, and this has allowed researchers to better examine empirical phenomenon. Some of the significant findings in the empirical public finance literature are discussed below, although because of the great breadth of the empirical literature, it is not possible to highlight every important area of empirical research.

Public finance economists have exploited changes in tax policy and individual and household-level data, such as the public use tax model files generated by the Internal Revenue Service's Statistics of Income Division, to measure how work incentives are distorted by tax policies. Empirical studies typically have found that the overall number of hours worked by men is not very responsive to changes in income tax rates. The high degree of inelasticity of the male labor supply implies that the burden of the income tax falls squarely on male workers, rather than employers, and that small increases in income tax rates are unlikely to create large deadweight losses from distortion of the incentives to work.

Even though, overall, the policy of income taxation may have a negligible effect on the aggregate supply of labor, income tax policy may significantly influence the labor supplies of specific groups of individuals and in specific markets. The labor supply of women, particularly married women, has been found to be much more elastic than the labor supply of men. Therefore, increases in income tax rates may cause significant numbers of married women to focus on household production instead of working in the workplace. Some studies have found that high marginal tax rates may induce some workers to leave legitimate occupations for work in the underground economy to avoid paying taxes. Also, the income tax policy pursued by a particular state may affect the labor supply of that state via migration. Some evidence suggests that workers tend to migrate from states with high income tax rates to states with low income tax rates.

Besides influencing labor supply, tax policy affects many other types of economic activity. Of particular interest is the effect of tax policy on savings decisions. Some economists have argued that lowering the tax rate on capital income increases savings because it increases the after-tax rate of return on investing. However, the theoretical and empirical evidence of the effect of tax rates on capital income and savings has been ambiguous. For example, if someone is saving with the explicit goal of accumulating $50,000 in 3 years for a down payment on a house, a reduction in capital income taxes will lower the amount this person needs to save now, because he or she can achieve the savings goal by saving less. Along these lines, more specific literature has examined how changes in tax policy that have given preferential treatment to retirement savings through contributions to IRA, 401(k), and other retirement plans have influenced aggregate retirement savings. Although some empirical evidence has indicated that this preferential tax treatment has had a positive influence on retirement savings, additional empirical analyses are still needed to develop a clearer understanding of this relationship.

Over the past 50 years, the role of the government in the economy has greatly expanded in terms of scope and magnitude, particularly regarding government spending on nondefense items. Nondefense spending by the federal and state governments in real 2007 dollars has increased almost fivefold between 1960 and 2007, from $686 billion in 1960 to $3,256 billion in 2007.[9] Spending by local governments has similarly swelled over this time period. This large expansion of the government has motivated public finance economists to examine the empirical effects of spending on a wide range of government programs.

Social insurance programs have been the recipients of much of the growth in government spending. A vast number of empirical studies have examined the economic issues for a host of social insurance programs, including Social Security; government health care programs for the elderly and poor, such as Medicare and Medicaid; unemployment insurance; worker's compensation; programs that provide subsidies to the poor, such as Temporary Assistance to Needy Families (TANF) and the Earned Income Tax Credit (EITC); and many other programs. Empirical research has found that social security programs in the United States and in other countries have reduced aggregate savings and produced early retirements. Studies have also examined the general equilibrium effects of Social Security reform, which would switch the current pay-as-you-go system to a system at least in part based on investment accounts similar to private retirement savings accounts. Many empirical studies on Medicaid and Medicare have focused on issues of adverse selection and moral hazard. Other studies have examined the potential fiscal impacts of continued growth in spending on Medicaid and Medicare.

The economic effects of government antipoverty programs have been the focus of many empirical studies. TANF is the primary welfare program in America today. In 1997, it replaced its predecessor, Aid to Families with Dependent Children (AFDC). In contrast to AFDC, TANF has strict time limits for assistance and places a strong emphasis on vocational training, community service, and the provision of childcare. Although more empirical work is still needed to sort out all of the economic impacts of TANF, one finding without controversy is that the switch to TANF has undoubtedly caused a large reduction in welfare caseloads. In terms of enrollment, EITC is the largest

government antipoverty program. EITC supplements the earnings of low-wage workers with a tax credit. This program may reduce the total hours worked of some workers and cause others to switch to lower-paying but more enjoyable occupations so that they can qualify for the tax credits. Notwithstanding these negative effects on labor supply, many empirical studies have found that EITC has increased the overall rate of labor force participation, particularly among single parents, by increasing the marginal benefit of working.

The structure of many government programs, such as public education, involves a complex interrelationship among the several tiers of government: federal, state, and local. Fiscal federalism is the subfield of public finance that examines the economic effects of having public sector activities and finances divided among the different tiers of government. A central question addressed by fiscal federalism is what aspects of the public sector are best centralized and what functions are best delegated to lower tiers of government.

Many empirical studies related to fiscal federalism have examined the economic effects of intergovernmental grants. The federal government gives grants to state and local governments for a variety of governmental programs, including those for public education, welfare, and public roads and highways. Many types of earmarks, such as the earmarking of state lottery revenues for educational funding, conceptually have the same economic effects as grants. According to standard economic theory, a lump-sum amount of intergovernmental aid given to a lower tier of government for an activity should have the same effect on spending for this activity as an increase in private incomes (Bradford & Oates, 1971). For example, a federal grant increase of $50 per taxpayer given to states for education should, theoretically, increase educational spending in the same manner as an earnings increase of $50 per taxpayer per year.[10] Therefore, grants and earmarks should have a minimal impact on total spending. Many empirical studies, however, have rejected the fungibility of intergovernmental aid, finding instead that the money given to lower tiers of government "sticks where it hits." As a result, this phenomenon is referred to as the *flypaper effect,* a term attributed to Arthur Okun. The findings in the empirical literature, however, are mixed. Some recent studies have found that local and state funding is substantially cut in response to funding from federal grants, particularly after time has elapsed since the federal aid was first received.

The Tiebout model outlined previously has spawned a rich empirical literature. Many empirical studies have built on Tiebout's work, using people's processes in choosing where to live to estimate the demands for local public goods, such as public education, and to measure how property values reflect local amenities and taxes. Empirical research has also used the Tiebout model as a foundation for explaining that the prevalence of restrictive zoning laws in many urban communities is a device to prevent free riding: buying inexpensive homes with low property tax assessments in areas where most residents own expensive homes with high property tax assessments. Additionally, the Tiebout model has influenced the fiscal federalism literature; many empirical studies have examined the proper roles of the different tiers of government.

## Public Policy Implications

It is not surprising that public finance has a wide array of implications for public policy, given that the focus of public finance is on the government's role in the economy and the economic effects of its actions. Much of the theoretical and empirical public finance literature has direct relevance for issues of public policy. Public finance economists work in all levels of government. They carry out a variety of important administrative functions and provide crucial advice and analyses to government decision makers. Public finance economists perform important public policy roles at all federal government agencies, including the Joint Committee of Taxation, the Congressional Budget Office, the United States Department of Agriculture, the Department of Defense, and the Department of Homeland Security. All state governments, and many county and city governments, also employ public finance economists, largely for their advice and analyses on issues of public policy. Even the public finance economists that are not directly employed by the government, working instead at universities and public policy institutions, directly influence debates of public policy through their research, writings, and consulting activities.

In their endeavors, public finance economists are influencing how to best address the key public policy issues of today. Here are just a few of the questions in the current debates over public policy that public finance economists are now trying to answer: What is the best way to keep Social Security sustainable without inducing severe distortions on work incentives? How should the federal government expand the coverage of heath care? How can federal and state governments best support local governments in their efforts to improve underperforming public schools? Should a state government lower corporate taxes to induce new businesses to relocate to their state?

Of course, public policy makers do not always heed the advice of public finance economists, nor do they always properly take into account the findings of economists when deciding on a course of action. In other situations, public policy makers receive conflicting advice from public finance economists, a result of vehement disagreement among economists over the proper course of action for issues of public policy. Nonetheless, the large number of public finance economists actively working with elected officials and other policy makers indicates that the field of public finance often does effectively influence public policy.

## Future Directions

In recent years, empirical research has flourished in the field of public finance. Every indication suggests that empirical research in public finance will continue to thrive in the near future. Advances in computer technology and econometric software will likely continue. New rich data sets will be created. Additionally, many new topics ripe for empirical analysis will surely arise. Governments will continue to tinker with tax policy. Reform and expansion of government social security programs will remain hotly debated topics. And the multiple tiers of government will continue to work together for providing many government programs, such as public education.

Many issues related to globalization will likely receive a lot of attention from public finance economists in the near future. By lowering the costs of capital and labor mobility across international borders, globalization has significant implications for many aspects of public finance, particularly for tax policy. Historically, some actors and rock stars have moved from one country to another, seeking lower taxes. Globalization has made this a viable option for a much larger segment of the world population. Internationally, capital is even more mobile than labor. Tax competition to induce businesses and residents to locate somewhere is no longer merely a local issue but an international one. Future research topics involving globalization will likely include how changes in tax policy induce international movements of labor and capital and how countries coordinate tax policy with each other.

The international financial crisis and federal financial bailout approved in the fall of 2008 will surely provide fruitful opportunities for future research in public finance. Many public finance issues are directly related to the federal bailouts. The federal financial bailout has important ramifications for issues related to tax burden and tax incidence, particularly if taxpayers are not fully reimbursed by the bailed-out industries. The federal bailout also involves many short- and long-run budgetary concerns. Another implication of the federal bailout is further expansion of the role of the government. Bailouts are already under way for the automobile industry. Will bailouts for other industries follow? Public finance economists will certainly examine the effectiveness of cooperation between the private sector and public sector officials.

## Conclusion

Public finance is the field of economics concerned with the role of the government in the economy and the economic consequences of the government's actions. Both positive and normative economic issues are discussed and debated in the public finance literature. Normative issues receive more attention in the public finance literature than in many other economic fields. Public finance theories examine taxation and government spending. Many theoretical advances have been made in public finance over the past 50 years. Theories of taxation examine tax incidence, economic efficiency costs of taxes, and the optimal design of systems of taxation. Issues of income inequality are sometimes taken into account by theoretical studies of the optimal design of taxation. The development of general equilibrium models has made important contributions to our theoretical understanding of taxation. Theories of government spending fall under two broad categories. First, some theories of government spending focus on the proper role of the government, examining the circumstances in which society is made better by government intervention in the economy. Second, some theories of government spending are more interested in why the government does what it does. In particular, the subfield of public choice is interested in why governments behave as they do. More recently, the empirical public finance literature has made great strides that have significantly added to our understanding of public finance. The empirical public finance literature has flourished, largely because of advances in computer technology, econometrics, and the creation of data sets with information at the household or individual level. A vast empirical literature now covers virtually every facet of taxation and government spending.

In the coming years, public finance economists will likely continue to advance our knowledge of the field and play an important role in influencing key debates over public policy issues. Globalization and the federal financial bailout that began in the fall of 2008 are two topics that surely will receive a lot of attention from public finance economists in the near future. It is to be hoped that public policy makers will rely on the advice and analyses offered by economists, many of whom specialize in public finance, to make wise decisions about the role of the government in the economy and to get us through the difficult economic times we now are experiencing.

## Notes

1. Sometimes public finance is defined more narrowly than public economics, with an emphasis primarily on issues of taxation and government expenditures.

2. The relative importance of normative issues in public finance is evident by the title of the popular textbook *Public Finance: A Normative Theory* by Richard W. Tresch (2002), which examines government economic policy.

3. Musgrave (1985) provides an excellent discussion of the history of fiscal doctrine.

4. See Samuelson (1954).

5. Because the supply curve with the tax is parallel to the original supply curve, the tax is constant per unit. Many taxes, such as sales taxes, are levied as a percentage of the products' prices. Nonetheless, the main economic findings are the same in either case.

6. This result assumes there are no negative externalities associated with the consumption of gasoline, particularly in relation to environmental degradation. Assuming there are negative externalities associated with gasoline consumption, then imposing a tax on gasoline could actually improve economic efficiency. See the discussion on Pigouvian taxes.

7. The general equilibrium models that were first used to examine incidence and deadweight loss of taxes were adopted from the economic literature on international trade. Harry Johnson (1974), an important figure in the early development of general equilibrium models for issues of international trade, provides a nice discussion of simple general equilibrium models in his *Two-Sector Model of General Equilibrium.*

8. The closed economy case is also useful for thinking about the situation in which many countries raise or lower their tax rates at about the same time.

9. Budget totals are calculated using the budgetary figures given by the Office of Management and Budget (1960, 2007).

10. Because education is a normal good, an increase in incomes would result in additional spending on education.

## References and Further Readings


Auerbach, A. J., & Hines, J. R. (2002). Taxation and economic efficiency. In A. J. Auerbach & M. Feldstein (Eds.), *Handbook of public economics* (pp. 1347–1421). New York: Elsevier.

Bradford, D. F., & Oates, W. E. (1971). The analysis of revenue sharing in a new approach to collective fiscal decisions. *Quarterly Journal of Economics, 85,* 416–439.

Buchanan, J. M., & Tullock, G. (1962). *The calculus of consent: Logical foundations of constitutional democracy.* Ann Arbor: University of Michigan Press.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics, 3,* 1–44.

Diamond, P. A., & Mirrlees, J. A. (1971a). Optimal taxation and public production 1: Production efficiency. *American Economic Review, 61,* 8–27.

Diamond, P. A., & Mirrlees, J. A. (1971b). Optimal taxation and public production 2: Tax rules. *American Economic Review, 61,* 261–278.

Feldstein, M. (2000). The transformation of public economics research: 1970–2000. In A. J. Auerbach & M. Feldstein (Eds.), *Handbook of public economics* (pp. xxvii–xxxiii). New York: Elsevier.

Gruber, J. (2007). *Public finance and public policy.* New York: Worth.

Harberger, A. C. (1962). The incidence of the corporation income tax. *Journal of Political Economy, 70,* 215–250.

Harberger, A. C. (2008). The incidence of the corporation income tax revisited. *National Tax Journal, 61,* 303–312.

Johnson, H. G. (1974). *Two-sector model of general equilibrium.* Portland, OR: Book News, Inc.

Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. *Review of Economic Studies, 38,* 175–208.

Musgrave, R. A. (1959). *The theory of public finance.* New York: McGraw-Hill.

Musgrave, R. A. (1985). A brief history of fiscal doctrine. In A. J. Auerbach & M. Feldstein (Eds.), *Handbook of public economics* (pp. 1–54). New York: Elsevier North-Holland.

Office of Management and Budget. (1960). *Budget of the United States.* Washington, DC: Government Printing Office.

Office of Management and Budget. (2007). *Budget of the United States.* Washington, DC: Government Printing Office.

Pigou, A. C. (1947). *A study in public finance.* London: Macmillan.

Ramsey, F. P. (1927). A contribution to the theory of taxation. *Economic Journal, 37,* 47–61.

Rosen, H., & Gayer, T. (2007). *Public finance.* New York: McGraw-Hill.

Samuelson, P. A. (1954). The pure theory of public expenditure. *Review of Economics and Statistics, 36,* 387–389.

Smith, A. (1776). *An inquiry into the nature and causes of the wealth of nations.* London: Methuen & Sons.

Tiebout, C. (1956). A pure theory of local expenditures. *Journal of Political Economy, 64,* 416–424.

Tresch, R. W. (2002). *Public finance: A normative theory.* San Diego, CA: Academic Press.

Wise, D. A. (Ed.). (1998). *Inquires in the economics of aging* (National Bureau of Economic Research Project Hardcover). Chicago: University of Chicago Press.

# 26

# Regulatory Economics

Christopher S. Decker
*University of Nebraska at Omaha*

Economic theory establishes that in a free and unfettered market, characterized by intense competition between many well-informed buyers and sellers, where resource mobility is possible, a socially efficient allocation of an economy's resources is generated. Characteristic of perfectly competitive markets, this social efficiency is measured as the maximization of total market surplus—that is, the sum of consumers' and producers' surpluses is maximized at a market-determined equilibrium price. For the individual firm, the consequence of this intensely competitive market is that price for the seller is determined within the market, and the only profit-maximizing mechanism available to the firm is to set production levels such that profits are maximized. This occurs when the firm's marginal cost of production equals the market price. Any deviation from this price will generate a reduction in market surplus, commonly referred to as a *deadweight loss.*

When a deadweight loss occurs, a market failure is said to result. This, then, becomes an economic justification for governmental intervention in the market place. History is replete with instances of government imposing restrictions on market outcomes and business behavior. For instance, for most of the nineteenth century, the government of the United States levied significant tariffs on imported goods in an effort to protect certain fledgling domestic industries from international competitive forces. Although there are many rationales, both noneconomic and economic, for governmental intervention in the marketplace, the focus of this chapter is specifically on economic rationales for governmental intervention.

In the United States (the primary focus in this chapter), there are two primary methods that government has adopted to impact business behavior and market outcomes. The first is referred to as *competition policy,* or *antitrust policy.* Competition policy tries to promote economic efficiency through rules, or codes of conduct, for firm behavior that are designed to restore the market to a competitive, or near competitive, outcome. It is not so much that monopolies are deemed illegal under competition policy, but rather it is more the attempt by a firm or group of firms to monopolize an otherwise competitive market that tends to trigger a government response. In many respects, competition policy tends to be reactive, in that it is prompted by legal challenges to a firm's observed behavior designed to garner market power when competition would otherwise exist. Competition policy is a vast field in economics and is beyond the scope of this chapter. This chapter focuses primarily on *direct regulation.*

Direct regulation is more preemptive in nature. Rules for market outcomes are established by the government. Prices are set, capital investments are subject to oversight, market entry is limited, and often certain types of production processes are mandated. Such regulation tends to override market outcomes that would have arisen without any such rules. From the perspective of most economists, who would generally prefer that the market allocate resources, the fact that regulation substitutes for market outcomes can be justified only if the market is indeed subject to failure.

After a brief review of the regulatory history of the United States, the economic reasons for regulation as well as common types of regulation are reviewed. However, because the economic prescriptions for market failures have not been followed closely in practice, a more formal review of the incentives of regulators to regulate is considered. Finally, with regulatory incentives more clearly understood, a review of deregulation is offered with an eye toward future industry regulatory efforts.

## A Brief Regulatory History of the United States

There are essentially four major periods in the regulatory history of the United States. The first occurred in the late nineteenth century. Article I, Section 8 of the U.S. Constitution grants the U.S. Congress the power to regulate commerce among and between states. This power is tempered by the Fifth Amendment to the Constitution, which asserts that the federal government shall not deprive any person the right to life, liberty, and property without due process. The first major test of government's authority to intervene in the marketplace was *Munn v. Illinois* in 1877. In a landmark decision, the U.S. Supreme Court upheld an Illinois law that regulated the prices that grain elevator operators could charge farmers. This case ultimately led to the passage of the Interstate Commerce Act in 1887 and the establishment of the Interstate Commerce Commission (ICC), which was initially set up to regulate the railroad industry.

The second major period was the progressive era and the New Deal, spanning the early twentieth century through the end of the 1930s. Regulation during this period was expanded dramatically. Industries such as electric utilities, telecommunications, trucking, air transportation, and financial securities were all subject to substantial regulatory rules and oversight. For instance, the Securities and Exchange Commission was established in 1934, and the Federal Energy Regulatory Commission was formed in 1930. Much of this regulation was focused on specific industries where prices, entry, capital investments, and quality of product and service were controlled.

The third major wave of regulation was in the 1960s and early 1970s. Rather than focusing on specific market outcomes such as prices and entry, these regulations were set up to address problems associated with product safety and the environment. For instance, the Consumer Product Safety Commission was established in 1972, and the U.S. Environmental Protection Agency was formed in 1970.

The fourth major period started in the mid-1970s and lasted through the early 1990s. During this period, many of the regulatory structures established in the 1930s were reversed, leading to substantial deregulation. Air transportation, rail and trucking industries, telecommunications, and financial markets were largely deregulated during this period. As we enter the twenty-first century, this last period of deregulation is prompting some reconsideration, at least in certain sectors of the economy. Addressing the motivations for both regulation and deregulation is clearly in order.

## Market Failure

As stated above, a market failure exists when the market, left to it its own devices, fails to efficiently allocate an economy's resources. Under such circumstances, market prices for goods and services will either be too high or too low, and production will either be too low or too high. There are several sources of market failures. For instance, a firm that enjoys limited competition, such as a monopolist, is in a position to select a profit-maximizing price in excess of the marginal cost of production. The general result is higher price, lower output, and the generation of a deadweight loss.

Another common form of market failure results from the presence of externalities in the market. An externality is a cost (or benefit) that is created as a result of market transactions that do not accrue to either producers or consumers. The classic example of an externality that generates a cost is pollution. The production of paper generates air and water pollution, thus degrading the surrounding natural environment. Residents located near the paper plant may not be employed by the paper plant and may not purchase any paper products produced by that plant. Yet they bear the cost of that pollution through potential respiratory problems and reduced quality of life. If the paper company does not compensate for this harm done, its marginal cost of production is lower than the social marginal cost of production (i.e., the cost that includes the harm done to the local residents). As a result, production levels will be too high and paper prices too low.

There are other types of market failure as well, the root cause of which has much to do with the very nature of the good (or service). Goods (or services) can be categorized based on their excludability and rivalry characteristics. A good is considered nonrival if consumption of the good by one person does not preclude other people from also deriving benefit from that good. Examples include radio signals, stadium lighting, an uncongested road, and national defense. A good is considered nonexclusionary if, once provided, no one can be prevented from consuming the good. An open access road for which there is no toll would be an example, as would, again, radio signals, stadium lighting, and national defense.

Generally speaking, markets will efficiently allocate resources in the production of goods that are rival and exclusionary, provided there is a sufficient amount of competition. For such goods, prices can be established, but no one can enjoy benefits without purchasing the good. Hence, firms will gain in such transactions. Goods that lack either excludability and/or rivalry will generate a market failure. For instance, common property resources are goods that are rival but nonexclusionary. A free and unfettered market cannot exclude access to the resource, but each additional user of that resource reduces the benefit other individual users derive from the resource. The classic example is an ocean fishery, whereby open access to the ocean prompts too much fishing activity, resulting in dramatic reductions in fish populations and potential resource extinction (the so-called tragedy of the commons). Hence, nonexcludability generates a market externality that would justify governmental intervention.

Goods that are both nonrival and nonexclusionary are referred to as *public goods.* Given these characteristics, potential suppliers of such goods (or services) would have little incentive to actually supply such markets because there is no way for a supplier to establish a price to charge consumers to secure a profitable return on investment. The result is a severe undersupply of the market, thus justifying some governmental intervention.

## Natural Monopolies

Both common property and public goods, as well as externalities, are subject to regulation. Such regulation is often termed *social regulation.* Because these types of conditions are addressed in other chapters in this volume, they are not addressed in detail here. However, goods that are nonrival and excludable dominate this chapter because they tend to generate monopoly markets where regulatory intervention is warranted on efficiency grounds. Markets where goods are nonrival and excludable are generally characterized as *natural monopolies.* Examples include electricity generation and distribution, telecommunications, cable television services, and natural gas pipelines. These goods are excludable in that access requires paying a price. However, these goods tend to be nonrival in that one additional consumer does not impact other consumers' utility. For instance, a new cable television subscriber pays a fee for the service, but that additional subscriber does not diminish other subscribers' services.

A hallmark characteristic of these markets is that there tends to be a substantial capital (fixed) cost component to supplying the good or service, but the marginal cost of supplying an additional unit tends to be relatively small. Consider an electric utility. There is a substantial capital cost associated with the construction and maintenance of generation and distribution equipment. However, once in place, the marginal cost of generating an additional kilowatt-hour of electricity is relatively small. Many goods and services in the information technology industry exhibit similar characteristics. For example, there is significant up-front investment in the development of new computer software. However, once developed, the cost of duplication and distribution to end consumers is relatively small.

A monopoly is considered natural if one firm can supply a good or a set of goods at a lower cost than can two or more firms. A key characteristic of such costs is called subadditivity. Costs are said to be subadditive if the following condition is met:

$$C(q_1) + C(q_2) + \ldots + C(q_n) > C(Q), \qquad (1)$$

where $Q = \sum_{i}^{n} q_i$.

This simply states the cost associated with producing at a total level of production, *C(Q),* by one supplier is lower than costs associated with producing smaller levels of production by many suppliers. The implication is that the average cost of production is declining over the relevant range of production. This is depicted in Figure 26.1.

Subadditivity is likely to arise under at least two circumstances. The first is in the presence of economies of scale where it is simply cheaper on average to produce larger levels of output. These scale benefits are likely to arise in industries such as electricity generation and distribution, where there is a substantial capital cost component. Spreading these large, fixed costs over larger levels of production reduces average total cost.

A second circumstance is in the presence of economies of scope (although it is possible for there to be diseconomies of scope and subadditivity). Economies of scope and subadditivity arise if a firm is producing multiple goods. If economies of scope exist, the fixed costs of production are larger than any increase in variable cost caused by joint production. Hence, it is cost effective to avoid the duplication of fixed costs associated with operating multiple plants and simply produce multiple goods from a single plant. Placed in the context of competition, then, production by multiple producers will duplicate fixed costs. Thus, if there are economies of scope, having a single firm producing output will be cost efficient.

While a single supplier is cost efficient in a natural monopoly, there is still an allocative efficiency problem if the monopolist is left to his or her own devises. Refer to Figure 26.1. Profit maximization results in a level of $Q_m$ and a price of $P_m$, generating deadweight loss equal to Area *abc.* Because promoting competition through entry will generate cost inefficiencies, some direct regulation is justified.

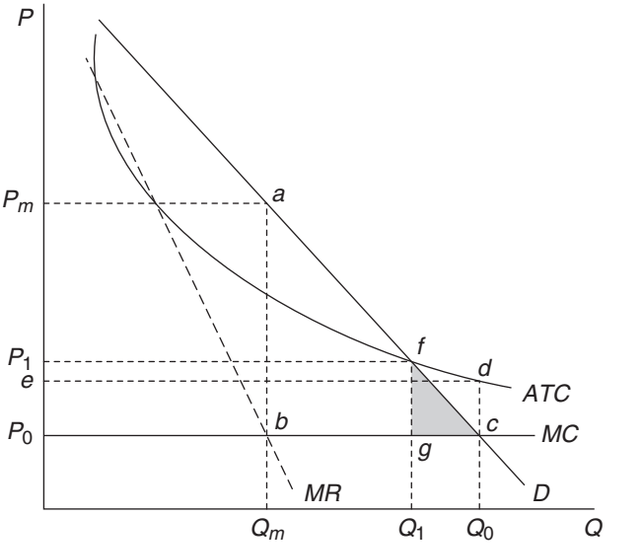


**Figure 26.1** Natural Monopoly

### Theoretical Solutions to the Natural Monopoly Problem

*Linear Marginal Cost Pricing*

The most obvious solution to the allocation problem is to regulate the price to equal marginal cost. In the figure, this would amount to setting price at $P_0$. This would result in increasing output to $Q_0$ and eliminating the deadweight loss. However, note that this will result in economic losses for the firm because at $P_0$ and $Q_0$, $P_0 < ATC$. Under such regulation, then, there would be no incentive for the firm to produce. What is necessary, then, is for the regulator to subsidize these losses by the amount $cdeP_0$, such that the firm earns zero economic profit (often referred to as a *fair rate of return*).

There are several problems with this solution. First, the regulator will likely not know the marginal cost, and the firm certainly has little incentive to provide accurate information to the regulator. Also, if the firm knows losses will be subsidized, there is less incentive for the firm to control production costs. Finally, even if all relevant cost information were known, there is an argument against this regulation on distributional grounds. Because the subsidy must come from public sources, it is not clear whether nonconsumers of this good being sold in this market should be required to subsidize producers based on economic efficiency arguments alone.

*Linear Average Cost Pricing*

In recognition that marginal cost pricing is not sustainable without subsidization, the regulator can default to average cost pricing where the price is set at $P_1$. At this price, $Q_1$ units of output are produced and the firm earns zero economic profit, another characteristic of (long-run) perfect competition. Note however, that this is a type of second-best solution to the natural monopoly problem in that there is still a deadweight loss of *cfg*.

## Government Regulation in Practice

The above discussion is largely theoretical. Implementation of regulation designed to mimic these outcomes has been the subject of substantial attention. Many means of correcting these allocation problems have been adopted. In this section, a few such policies are addressed.

### Public Ownership and Operation

Perhaps the first thought on how to correct market failure is to remove the market and allow government to supply the good. The classic example in the United States is the U.S. Postal Service. U.S. ports are also largely government owned. While relatively limited in the Unites States, public ownership and operation is more widespread in other largely market-oriented economies. In Australia, for instance, ports, electric utilities, telecommunications, natural gas distribution, and railways are nearly entirely owned and operated by the state. With state ownership, the price can presumably be set to maximize total market surplus.

Much has been written and debated about the relative merits of public versus private provision of goods and services. Many studies, such as those by Louis De Alessi (1980) and Anthony Boardman and Aidan Vining (1989), find public enterprises are less responsive to market changes, operate less efficiently, and adopt cost-reducing technologies more slowly. Other studies, such as that by William Hausman and John Neufeld (1991), find evidence in the electric utility sector (prior to rate-of-return regulation, discussed below) that publicly owned firms were more efficient than those that were privately owned. This is an issue that has been debated for years and will likely continue in the foreseeable future.

### Franchise Bidding

Harold Demsetz (1968) questions the very need to regulate utilities or other natural monopolies on efficiency grounds. His argument is that even though feasible competition in the supply of such goods is not possible, one can introduce competition by allowing firms the opportunity to bid for the right to supply the market. In essence, government auctions off the right to supply the market to the least-cost bidder. As long as there is sufficient competition—that is, a sufficient number of least-cost bidders—in the bidding process, the winning bid would be the one where the price charged to consumers would be just sufficient to ensure zero economic profit—that is, $P = P_1$ in Figure 26.1. Although this price does not eliminate the deadweight loss, franchise bidding imposes few informational requirements on a government agency. Franchise bidding, while not widely adopted in the electric utility industry, has been used rather extensively in supplying cable television services in the United States.

Although there appear to be benefits to franchise bidding, there are some potential drawbacks. For instance, if competition is over price alone, product or service quality might suffer. Offering a good (or service) at least cost may involve poorer quality inputs, little attention to manufacturing quality, poor customer services, and so forth. The duration of the franchise contract is an issue as well. If a firm is granted a franchise for a relatively long term, there must be some contractual delineations for rate changes as input costs, consumer demand, and technologies change. This may require the existence of a regulatory body to oversee such changes. If the contract is relatively short in duration, then the franchisee may, toward the end of the contract, skimp on maintenance and underinvest in new capital, leaving real problems for the next franchise owner.

## Rate-of-Return Regulation

Rate-of-return regulation is quite common and has been mostly directed at the electric utility industry. In essence, government (typically state-level government) sets up regulatory commissions. These commissions regulate the prices that the firm can charge its customers so that the firm earns a normal profit. However, it does so in a rather indirect way by regulating the rate of return the firm can earn on capital investments. For instance, maintenance and expansion of electricity generation plants and transmission lines require capital expenditures by the firm. Ideally, the firm would pass some of this cost on to consumers in the form of higher prices. To ensure that the needed price increase allows the firm only a normal rate of return on that investment (akin to what linear average cost pricing would justify), a rate hearing is convened and a new rate is agreed on. In effect, the allowed rate of return on the capital investment (i.e., the price of capital) is set. In turn, the price the firm charges consumers is then adjusted. Although this is an indirect means of attempting to achieve linear average cost pricing, it has become the most common regulatory form in the United States.

Several issues arise here as well. First, it may be that the regulatory commission allows a higher than fair rate of return if information is not complete enough to ensure an accurate cost measure. Another issue that arises is regulatory lag. Once a new fair rate of return on investment is agreed on, capital market prices (i.e., interest rates on purchased capital) can change. Before a new rate hearing can be set, there may be a significant period of time when the firm can enjoy a rate of return on its investment in excess of the fair rate of return.

The potential for the market rate of return on capital investments to differ from the regulated rate of return has caused economists to question the efficiency of rate-of-return regulation. In a seminal article, Harvey Averch and Leland Johnson (1962) show that rate-of-return regulation could lead to overcapitalization in production. Their basic model involves the regulated firm maximizing profits by choosing a level of capital, $K$, and labor, $L$:

$$\max_{K,L} \Pi = R(K,L) - wL - rK, \tag{2}$$

subject to the following regulatory constraint:

$$\frac{R(K,L) - wL}{K} = s, \tag{3}$$

where $R(K,L)$ is the firm's revenue function, $w$ is the wage rate, $r$ is the unit cost of capital, and $s$ is the allowed rate of return on $K$. Note that if $s < r$, then the firm will not produce. So we consider the case where $s \geq r$. Optimization yields:

$$\frac{MPK}{MPL} = \frac{r - \alpha}{w}, \tag{4}$$

where $MPK$ and $MPL$ are the marginal products of capital and labor, respectively, and

$$\alpha = \frac{\lambda(s - r)}{1 - \lambda}, \tag{5}$$

where $\lambda$ is the Lagrange multiplier whose value lies between zero and 1. Note that if $s = r$, then $\alpha = 0$ and the result would be the standard long-run condition for optimal input mix presented in intermediate microeconomics courses:

$$\frac{MPK}{MPL} = \frac{r}{w}. \tag{6}$$

In effect, the regulator has ensured not only a normal economic profit but also optimal input mix of $K$ and $L$. However, if $s > r$, then $\alpha > 0$. Under this condition, after some algebra, we find that rate-of-return regulation causes the following result:

$$\frac{MPK}{r} > \frac{MPL}{w}. \tag{7}$$

The condition of Equation 7 states, under rate-of-return regulation, that at the prevailing market prices for $K$ and $L$, the firm has an incentive to buy more capital. In effect, regulation has induced the firm to value capital at a higher rate than the market for that capital dictates, prompting additional purchases of $K$. This overcapitalization result prompted a series of empirical studies. Some have found support for the model, such as H. Craig Petersen (1975); others have not, such as William Boyes (1976).

The potential for resource misallocation under rate-of-return regulation can have other undesirable implications. Joseph Flynn and John Mayo (1988), for instance, find evidence that the regulated rate of return does have a significant effect on the incentives of electric utilities to invest in research and development. Indeed, they find that the more restrictive the regulated rate—that is, the lower $s$ is, the less research and development will be undertaken by regulated firms. In effect, the firm is responding to the regulatory constraint, rather than market forces, in seeking cost-efficient innovations to production and distribution. Given that rate hearings are costly and potentially infrequent, the potential implication of these results is that rate-of-return regulation can slow innovation if $s$ is low and not allowed to increase. In response to such possibilities, economists have developed incentive-based policies that allow regulated firms opportunities to respond to market conditions.

## Cross's Incentive Pricing Scheme

John Cross (1970) developed a simple pricing rule that would allow regulated firms to reap rewards for adopting cost-efficient innovations. His basic rule is

$$P = \beta + \alpha(ATC), \tag{8}$$

where $P$ is the regulated price, $ATC$ is the regulated firm's average total cost, $\beta$ is a base price below which actual prices may not fall, and $\alpha$ is a sharing rate coefficient bound between zero and 1. This sharing rate measures the percentage of $ATC$ that gets reflected in price. Recall that rate-of-return regulation ideally tries to set price equal to $ATC$ and hold it there until the next rate hearing. To see how Cross's rule works, suppose that $ATC$ is set at \$10 per unit, $\beta$ is \$5 per unit, and $\alpha$ is 0.5. Clearly, under Cross's rule, $P$ is \$10 per unit. Now suppose that the firm is able to reduce average costs to \$9 per unit. The new price $P$ is thus \$9.50. Note that consumers gain by enjoying a lower price. However, the firm also gains because average costs have fallen by \$1 per unit but price has fallen only 50¢ per unit. Hence, the firm gets to enjoy the benefits associated with its cost reduction efforts automatically.

### Price Cap Regulation

Cross's rule has not appeared to have realized much application. However, a related incentive-based regulatory scheme called *price cap regulation* has. Under this regulatory policy, a firm's price is set, or capped, at an initial regulatory hearing. After that, the regulated firm is allowed to adjust price at the rate of inflation minus the rate of productivity growth. So if inflation were running at 4% and productivity gains were averaging 1%, the firm could increase its price by 3%. Moreover, the firm is free to price at any level it chooses below the price cap. Like Cross's rule, allowing automatic price changes due to both market conditions and technology offers the firm incentives to seek productivity gains. Such regulation was adopted in the 1980s in the telecommunications industry. Alan Mathios and Robert Rogers (1989) study states with some type of price cap mechanism versus those without. They found that those states with such a price cap regulation had lower prices for long-distance service than those without.

There are many other incentive-based regulatory schemes that seem to offer results to superior rate-of-return regulation, but their adoption in policy circles is relatively slow. The natural question to ask is why. Perhaps a more formal review of the motivation for regulation is in order.

## Regulatory Behavior and Theories of Regulation

Irrespective of the type of policy adopted to regulate natural monopolies, one underlying motivation seems clear: There is a general recognition that, left to its own devices, the market will generate a higher price and a lower level of production than is socially desirable. The question, however, is whether regulators are in fact motivated to adopt policies that lower price to increase social welfare or are responding to other forces in the economy.

### Public Interest Theory of Regulation

The idea that regulators are motivated to maximize social welfare by lowering prices and reducing deadweight losses forms the foundation of a theory of regulation commonly referred to as the *public interest theory* of regulation.

The essential validity of this theory was challenged by several scholars, starting in the 1960s. In a very influential article, George Stigler and Claire Friedland (1962) study prices charged to consumers by electric utilities in states where some regulation existed and compared them with rates levied in states with no regulation. They find, contrary to what the public interest theory would predict, no statistical difference in prices between these groups of states. Although this prompted many follow-up empirical studies, some of which confirm Stilger and Friedland and some of which refute their findings, there is little doubt that this study prompted economists to reconsider the public interest theory of regulation. Indeed, in a detailed and often-cited review of regulation in the United States, Richard Posner (1974) concludes that there is little correlation between regulation and the correction of externalities or the curbing of monopoly power. Clearly, revising regulatory motivation was in order.

### Economic Theory of Regulation

Development of an alternative theory of regulatory behavior is rooted in four seminal studies: Mancur Olson (1965), Stigler (1971), Sam Peltzman (1976), and Gary Becker (1983). Stigler, building on Olson's work on the rationale for collective action, addresses these fundamental questions: Why is there regulation? How is it implemented? He then sets forth a series of general assumptions that would enable predictions as to which sectors of the economy would be subject to regulation and which would not. This analysis is formalized by Peltzman.

#### *Stilger and Peltzman Models*

The basic insight and structure of the theory is as follows. First, Stigler (1971) asserts that government is unique in its power to coerce agents in an economy to act or behave in certain ways. This basic resource—that is, the power to coerce—is in effect a tradable commodity. An interest group, such as an industry or a labor union, that can influence the government to employ its power of coercion to meet the desires of that group, stands to gain. Therefore, there is in fact a market for political influence, the supply of which is the power of coercion, and the demand for which is the potential benefits the coercion can offer an interested party. The mechanism through which political influence is exchanged can take many forms. Perhaps the most common are the securing of monetary campaign contributions and the securing of votes for

particular candidates. The outcome of the transaction may be varied as well. Indeed, regulation may be the primary avenue whereby an interest group can increase its economic gains by having the government construct rules and institutions that favor that group. By relying on market forces to dictate the structure and direction of regulation, predictions can be made regarding where regulation is likely to occur and how. In effect, any group that stands to benefit and is able to pay the price for such regulation will realize regulatory benefits.

*Types of benefits received by an interest group.* There are many types of benefits a group receiving favorable regulation can receive. One benefit is a direct subsidy from the government. The most famous example is the U.S. government land grants given to the transcontinental railroads in the 1860s. Another regulatory benefit might manifest itself in the form of an entry barrier allowing existing firms to enjoy secure market shares and profits. Since the 1930s, many large cities in the United States have limited the number of legally operating taxicabs by issuing taxi medallions (essentially, permits to operate a taxicab). Regulation can also increase the price of a substitute product for a particular industry. For instance, in the 1930s and 1940s, the ICC expanded its authority to regulate freight rates for trucking and barges, in effect increasing rates in these sectors. This benefited the rail industry, which otherwise might have faced substantial price competition from these alternative services. Finally, it is possible for regulation to favor one group of customers over another. This is a form of cross-subsidization whereby different prices are charged to different groups. For instance, it is generally accepted today that prior to deregulation of the telecommunications industry, because the rates for long-distance service were substantially higher than rates for local services, users of long-distance service were in effect subsidizing local service customers.

*Characteristics of groups that tend to garner regulatory favor.* The literature suggests at least five characteristics of groups that garner regulatory favor. The first is perhaps the most important. The size—that is, the number of members—of the interest group matters. At first glance, one might conclude that the larger the membership of the interest group, the more influence that group will wield. While this is in part the case, there is an important element to consider. In the exercise of political influence, which can be a very costly activity, there is serious potential for a free-rider effect to arise. If an interest group is sufficiently large, one member may rationally deduce that one additional player's marginal contribution to political influence is sufficiently small relative to the cost of participation and choose not to participate in the influence game, believing that the other members will engage. This is the essence of the free-rider effect—that is, attempting to secure benefits on the efforts of others without having to incur any costs. However, such a rationale is likely to be exercised by others in the group as well. As a result, there is likely to be a substantial underrepresentation of political influence by the interest group. A smaller group stands a better chance of engaging all members in political influence because it is easier to identify, and punish, free riders. Hence, in the regulatory influence realm, moderately sized interest groups are likely to have more success than very large groups or relatively small groups.

Second, groups with a strong commonality of interests tend to be more successful in regulatory influence. Many groups have divergent interests. Other groups have very well-defined interests and objectives. The voting population of the United States, for instance, constitutes a political interest group. This group clearly has a wide range of interests, and it is difficult to identify a given issue or objective that dominates this group's concerns. However, firms competing in the U.S. steel industry and facing the prospect of foreign competition due to relaxed trade quotas or lower trade tariffs have a strong common interest in hindering such legislation. As a result, they are likely to be more successful in achieving their objectives.

Third, those interest groups with a very large overall per-member benefit from a given regulatory outcome are going to have a greater incentive to engage in political influence, and they are likely to exert greater time and effort in obtaining their desired outcomes. Fourth, success often hinges on what is termed the *deep pockets effect.* Those interest groups that have greater financial resources to support political causes or candidates are likely to have greater success in the regulatory influence game. Finally, interest groups with complete information about the various ramifications of regulatory structures are more likely to succeed. The ultimate effects of many regulatory actions are often far from obvious. If a given regulatory outcome is not clearly understood by a particular interest group affected by the regulation, that group's interests will be underrepresented in the political process.

In short, then, those groups with an optimal number of highly impacted members with ready access to good information regarding regulatory outcome as well as sufficient resources available to influence regulators are likely to have success in influencing the structure and enforcement of regulations. The theory of economic regulation, in general, does not necessarily focus on an industry's ability to direct regulatory structures in its favor. Indeed, there are many interest groups that exert substantial influence on regulators. When it appears that regulation favors an industry, the term *capture theory of regulation* is often adopted.

Capture theory posits that legislators that design regulations favoring an industry are in effect captured by that industry. Many examples seem to fit well with the characteristics of this theory. Randal Rucker and Walter Thurman (1990) highlight the U.S. peanut market as an example. Since 1949, the federal government managed a program

called the U.S. Peanut Program, which had the effect, through international quota limits and other means, of limiting the number of farmers who could sell peanuts in the United States. The number of U.S. peanut farmers is relatively small, but they have enjoyed substantial profits from peanut sales. Although this program in effect kept peanut prices higher than a competitive market would generate (estimates suggest that the domestic price is about 50% higher than the world price), peanut demand is relatively broad based, and consumers, representing a very large constituency with diverse interests, are ill inclined to mobilize against a program that is estimated to cost each consumer only $1.23. Clearly, the capture theory of regulation appears to have been in operation in this market.

### *Becker's Model*

The Stigler (1971) and Peltzman (1976) models of economic regulation are essentially based on a regulator's objective of maximizing political support. Building on this basic structure, Becker (1983) focuses attention on competition between interest groups. In Becker's construct, there is a fixed amount of political influence, and interest groups compete for some portion, if not all, of this influence. Interest groups, in deciding how much political pressure to apply on the regulator, in the form of, say, direct lobbying or campaign contributions, must now consider not only the cost of such activity but also the level of political pressure their competitors are likely to exert. This approach introduces a game theoretic concept into the economic theory of regulation: the theory of best response. In Becker's model, one interest group's best response in increases in political pressure by a competing group is to itself increase its level of political pressure. Ultimately, this will lead to a political-pressure equilibrium, where political influence is, in effect, shared between groups. One interest group enjoys a greater share of political influence than another if the characteristics of that group are favorable to success (i.e., a optimal group size, sufficient resources, and complete information). Given the determination of relative influence as the outcome, however, Becker's approach focuses on relative effectiveness of interest group influence. Specifically, an interest group could succeed in gaining a substantial amount of political influence even if there is substantial free-riding behavior among its members. What is important now is that the interest group's free-riding problem is less severe than that of any competing interest group. It is the relative severity of free riding that leads to the relative distribution of political influence.

## The Deregulation Movement and Beyond

With a clear indication from academic economists, policy analysts, and general public perception that regulation was not only failing to generate social-welfare-enhancing outcomes but also appearing to favor the economic interests of regulated industry, a wave of deregulation efforts began in the mid-1970s. Although the long-term implication of mass deregulation is subject to considerable debate, there are a number of instances where evidence suggests that the immediate aftereffects of deregulation in certain sectors were quite positive.

One example is the airline industry in the United States. In 1938, the Civil Aeronautics Act was passed, creating the Civil Aeronautics Board, which effectively regulated price, routes, entry, capital investment, and scheduling of flights. However, in the intervening years, it became clear that regulation of this industry was not justified, on grounds that it exhibited characteristics of a natural monopoly. Indeed, most economists viewed the industry as a workably competitive one. Such a market is characterized by a sufficient number of firms to warrant potential competition, limited scale economies, mobile capital, and relatively frictionless entry and exit. The airline industry generally has these characteristics: There are a number of competitors, entry into viable markets is quite possible, and planes are highly mobile resources. Eventually, the industry was effectively deregulated in 1978. According to Alfred Kahn (1988), the result was that between 1976 and 1986, fares fell over 28%. Moreover, Steven Morrison and Clifford Winston (1986) estimate welfare gains to the industry of about $8 billion.

Another industry with a long history of regulation that was deregulated at this time was the railroad industry. As stated, the regulations of rates, entry, and capital investment began in 1887 with the passage of the Interstate Commerce Act. The Hepburn Act of 1906 gave the ICC the authority to set maximal rates, and the Transportation Act of 1920 gave it the authority to set minimal rates and regulate entry on rail routes. The reason for this regulation has been linked directly to the economic theory of regulation. In short, the existing railroad companies of the early twentieth century in effect wanted regulation to stabilize prices, secure market shares, and limit substitute services, such as trucking or barge transportation, from undercutting their business.

Over time, however, it became clear that regulation was preventing efficient pricing. In effect, this industry was also a working competitive one where competitive pricing was possible and allocatively efficient. In 1980, the Staggers Act was passed, effectively ending the ICC's control over the industry. The result, according to Kenneth Boyer (1987), was that freight rates fell over 12% between 1981 and 1986. Regulation was, in effect, artificially inflating prices.

## Conclusion: Is Reregulation in the Future?

This chapter offers a brief overview of the regulatory economics field. Given space limitations, a number of key subjects have been omitted. Interested readers are referred to W. Kip Viscusi, John Vernon, and Joseph Harrington

(2005) and David Kaserman and John Mayo (1995) for more in-depth study. That said, this chapter highlights the key literature in the field.

The challenges facing the U.S. economy at the beginning of the twenty-first century are quite substantial. Although this chapter has intentionally not focused on regulation of the financial markets, largely because many of the issues are macroeconomic and finance-oriented in nature, needing an entire chapter to effectively evaluate, it seems clear that dramatic deregulation of the financial markets in the 1990s may have gone too far. Some form of reregulation may be in order. Indeed, reregulation may need to be reconsidered in other sectors as well. Future policy recommendations coming from economists regarding such regulation are unknown. However, analysis is likely to build on some of the incentive regulations discussed in this chapter, with deference toward the economic theory of regulation, and focus on information asymmetries, another type of market failure that generates an inefficient price and output combination. Whether an industry is a natural monopoly or not, be it automobiles, electricity, or financial securities, sellers tend to have more information than buyers, and the incentives to exploit such informational advantages are quite powerful. Consumers could be induced to pay a higher price than a competitive market with full information would justify. Regulation is likely to focus on rules that require complete disclosure of such information. How such policies should look will require substantial scholarly attention. The regulatory economics is a fascinating subject indeed, and, with much work yet to be done, this field's future is bright.

## References and Further Readings


Averch, H., & Johnson, L. L. (1962). Behavior of the firm under regulatory constraint. *American Economic Review, 52,* 1023–1051.

Becker, G. S. (1983). A theory of competition among pressure groups for political influence. *Quarterly Journal of Economics, 98,* 371–400.

Boardman, A. E., & Vining, A. R. (1989). Ownership and performance in competitive environments: A comparison of private, mixed, and state-owned enterprises. *Journal of Law and Economics, 32,* 1–34.

Boyer, K. D. (1987). The costs of price regulation: Lessons from railroad deregulation. *RAND Journal of Economics, 18,* 408–416.

Boyes, W. J. (1976). An empirical examination of the Averch-Johnson effect. *Economic Inquiry, 14,* 25–35.

Cross, J. G. (1970). Incentive pricing and utility regulation. *Quarterly Journal of Economics, 84,* 236–253.

De Alessi, L. (1980). The economics of property rights: A review of the evidence. *Research in Law and Economics, 2,* 1–47.

Demsetz, H. (1968). Why regulate utilities? *Journal of Law and Economics, 11,* 55–65.

Flynn, J. W., & Mayo, J. W. (1988). The effects of regulation on research and development: Theory and evidence. *Journal of Business, 61,* 321–336.

Hausman, W. J., & Neufeld, J. L. (1991). Property rights versus public spirit: Ownership and efficiency of U.S. electric utilities prior to rate-of-return regulation. *Review of Economics and Statistics, 78,* 414–423.

Kahn, A. E. (1988). Surprises of airline deregulation. *American Economic Review, 78,* 316–322.

Kaserman, D. L., & Mayo, J. W.(1995). *Government and business: The economics of antitrust and regulation.* New York: Dryden Press.

Mathios, A. D., & Rogers, R. P. (1989). The impact of alternative forms of state regulation of AT&T on direct-dial, long-distance telephone rates. *RAND Journal of Economics, 20,* 437–453.

Morrison, S., & Winston, C. (1986). *The economic effects of airline deregulation.* Washington, DC: Brookings Institution.

Olson, M. (1965). *The logic of collective action.* Cambridge, MA: Harvard University Press.

Peltzman, S. (1976). Toward a more general theory of regulation. *Journal of Law and Economics, 19,* 211–240.

Petersen, H. C. (1975). An empirical test of regulatory effects. *Bell Journal of Economics, 6,* 111–126.

Posner, R. (1974). Theories of economic regulation. *Bell Journal of Economics and Management Science, 5,* 335–358.

Rucker, R. R., & Thurman, W. N. (1990). The economic effects of supply controls: A simple analytics of the U.S. peanut program. *Journal of Law and Economics, 33,* 483–515.

Stigler, G. J. (1971). The theory of economic regulation. *Bell Journal of Economics and Management Science, 2,* 3–21.

Stigler, G. J., & Friedland, C. (1962). What can regulators regulate? The case of electricity. *Journal of Law and Economics, 5,* 1–16.

Viscusi, W. K., Vernon, J. M., & Harrington, J. E., Jr. (2005). *The economics of regulation and antitrust* (5th ed.). Cambridge: MIT Press.

# 27

# Cost-Benefit Analysis

Mikael Svensson
*Örebro University, Sweden*

Infrastructure, education, environmental protection, and health care are examples of goods and services that in many circumstances are not produced by competitive private companies. Instead, decision making regarding investments and regulations is often made by politicians or public sector officials. For these decisions to be consistent, rational, and increase welfare, a systematic approach to evaluating policy proposals is necessary. Cost-benefit analysis is such a tool to guide decision making in evaluation of public projects and regulations. Cost-benefit analysis is a procedure where all the relevant consequences associated with a policy are converted into a monetary metric. In that sense, it can be thought of as a scale of balance, where the policy is said to increase welfare if the benefits outweigh the costs. Cost-benefit analysis of a proposed policy may be structured along the following lines:

1. *Identify the relevant population of the project.* For a cost-benefit analysis of a single individual or for a firm, this is not a problem. But in a societal cost-benefit analysis, we need to consider how to define society. A common approach is to consider the whole country as the relevant population. This is reasonable given that most public policies are financed at the national level. Another approach is to conduct a cost-benefit analysis specifying costs and benefits using different definitions of the relevant population—for example, including benefits and costs of a neighboring country in the analysis.

2. *Specify all the relevant benefits and costs associated with the policy.* The aim with a cost-benefit analysis is to include all relevant costs and benefits with a policy. Based on the definition of the population (Step 1), every aspect that individuals in this population count as a benefit or a cost should be included in the analysis. For some benefits and costs of a policy, this may be an easy task—for example, 2,000 hours of labor input are required next year to build a road. But there are also more difficult phases in this step of a cost-benefit analysis; some examples include (a) a new infrastructure investment may have ecological consequences that are difficult to estimate and (b) regulating speed limits in a major city may have beneficial health effects due to decreases in small hazardous particles. For the economist or analyst, this step often consists of asking the right questions and gathering the necessary information from the literature, professionals, or both.

3. *Translate all the benefits and costs into a monetary metric.* A cost-benefit analysis requires that the different consequences are expressed in an identical metric. For simplicity, we use a monetary metric ($ or €, etc.). Consider the example of an investment in a new road. The costs for the labor input can be valued by the wages (plus social fees, etc.), and the use of equipment may be estimated by the machine-hour cost. For benefits of, for example, increased road safety, decreased travel time, and decreased pollution level, there are no market prices to use; instead, weights and estimates are necessary to translate these benefits into a monetary metric. If this would not be possible with all relevant consequences, they should at least be included as qualitative terms in the evaluation. In the Valuation of Nonmarket Goods section, giving nonmarket goods a monetary estimate is discussed.

4. *If benefits and costs arise at different times, convert them into present value using an appropriate social discount rate.* Most people prefer a benefit today to a benefit in one year. There are several reasons for this, one being that no one knows for sure whether he or she will be alive in a year. Another reason may be that, over a longer

time horizon, people expect their incomes to increase in the future. If an extra dollar has a larger utility benefit to a less rich individual, that individual would prefer to consume when he or she has less money (today) rather than when he or she has more (e.g., in 5 years). Also, from an opportunity-cost approach, $100 that is not consumed today can be invested in a bond, and in a year from now, it may be worth $105, implying higher consumption in one year. In a cost-benefit analysis, economists therefore (generally) do not treat benefits and costs that occur at different times as equal; rather, they translate all benefits and costs into a present value. Choosing an appropriate social discount rate is, however, a complicated task and will often have major effects on the results of the evaluation. In the Discounting section, this is discussed in more detail.

5. *Compare the net present value of benefits and costs.* When the present value of benefits ($PV_B$) and costs ($PV_C$) has been calculated, what remains is to calculate the net present value (*NPV*). The policy is said to increase social welfare if the net present value is positive—that is, $PV_B - PV_C > 0$. It is also common to express the comparison as the benefit-cost ratio—that is, $PV_B / PV_C$, which gives the relative return of the investment. If the ratio is greater than 1, the policy increases social welfare.

6. *Perform a sensitivity analysis to see how uncertain the benefit-cost calculation may be and give a policy recommendation.* A cost-benefit analysis will have several uncertainties regarding the outcome. It is most often not reasonable to show only one point estimate of the evaluation. There are often uncertainties regarding both parameter values (such as the monetary estimates of, e.g., increased safety or environmental pollution) and more technical issues, such as the economic lifetime of a new road, which is the period during which it retains its function. These uncertainties need to be explicitly modeled. The final step in a cost-benefit analysis is to give a policy recommendation based on the result of the evaluation as well as the uncertainties associated with the result.

## What Do We Mean by Social Welfare?

The aim with a cost-benefit analysis is to evaluate the welfare effect of a policy. This requires a definition of what is meant by social welfare in an economic framework. The meaning of a welfare improvement is, in its most restricted view, formulated in the Pareto criterion. The Pareto criterion states that a policy that makes at least one individual better off without making any other individual worse off is a Pareto-efficient improvement and increases welfare. However, the Pareto criterion is generally useless as a definition for welfare improvements in a real-world application, because more or less all policies make at least someone worse off. It may also be criticized on ethical grounds; consider a vaccine that would save 1 million lives in sub-Saharan Africa but required a €1 tax on someone (a nonaltruistic individual) in Europe or the United States. According to the Pareto criterion, this policy could not be said to increase welfare, but this conclusion would violate the moral values of most individuals.

As a development to the Pareto criterion, the Kaldor-Hicks criterion was formulated, and it may be seen as the foundation of practical cost-benefit analysis. The Kaldor-Hicks criterion is less restrictive than the Pareto criterion and may be interpreted such that a policy is considered to increase welfare if the winners from a policy are made so much better off that they can fully (hypothetically) compensate the losers and still gain from the policy (potential Pareto improvement). This implies that every Pareto improvement is a Kaldor-Hicks improvement, but the reverse is not necessarily true. The compensation from the winners to the losers is a hypothetical test, and the compensation does not need to be enforced in reality (which distinguishes it from the Pareto criterion). In the example above, the vaccine to save 1 million lives would pass the Kaldor-Hicks criterion, if those who were saved would be hypothetically willing to compensate the European taxpayer with a payment of at least €1. Cost-benefit analysis is a test of the Kaldor-Hicks criterion, translating all the benefits and the costs into a monetary metric. The Kaldor-Hicks criterion also implies that we need to collect information only on aggregate benefits and costs of a policy; we do not need to bother ourselves determining which individuals are actually winning or losing from a policy.

## Valuing Benefits and Costs

Performing a cost-benefit analysis of a policy requires that all benefits and costs of the policy be summed up in monetary terms. There are two principal ways of measuring the sum of benefits in a cost-benefit analysis: willingness to pay (WTP) and willingness to accept (WTA). Both WTP and WTA are meant to value how much a certain policy is worth to an individual in monetary terms. The WTP of a policy may be characterized as an individual's maximum willingness to pay, such that he or she is indifferent to whether the policy is implemented—that is, if it is implemented, utility is the same after the policy as before the policy. The WTA of a policy may instead be characterized as the lowest monetary sum that an individual accepts instead of having the investment implemented—that is, utility is the same when receiving the money as it would have been if the investment had been implemented. WTP and WTA for a policy are expected to differ (WTP < WTA) because of the income effect. Robert D. Willig (1976) shows, using plausible assumptions, that WTP and WTA should not differ from each other by more than a few percentage points. But a lot of research has

documented that WTP and WTA usually differ a lot for the same policy, with WTA being significantly higher than WTP. Several hypotheses have been put forward trying to explain this discrepancy, one frequent hypothesis being the existence of an endowment effect (Kahneman, Knetsch, & Thaler, 1990). An endowment effect states that individuals value a good significantly more once they own it, which would create a gap between WTP and WTA for an identical good. In several circumstances, it is often more problematic to estimate WTA for a good, especially in valuation of nonmarket goods, implying that it is more common to use the concept of WTP in cost-benefit analyses.

The cost of a certain policy is based on the opportunity cost concept—that is, the value of the best alternative option that the resources could be devoted to instead. The opportunity cost is derived based on companies' marginal cost curve—that is, the cost that is associated with increasing output with one unit. This should not be equalized to accounting costs, which do not necessarily tell us the economic cost of an activity. As an example, going to a 2-year MBA program has some direct costs (accounting costs), such as tuition fees, books, and travel costs to attend lectures, but also indirect costs, such as the money that one could earn in a job if not attending the MBA program. The economic cost of an activity is the sum of direct and indirect costs.

## Valuation of Market Goods

To estimate benefits and costs of a policy, the natural starting point is to examine whether market prices exist that may be used. If the market is characterized by perfect competition, or a reasonable approximation of perfect competition, and no external effects exist, the market price can give us information about the willingness to pay for a marginal change of a good. For example, if we need to use production factor $X$ for a certain investment and our demand for $X$ will not affect the market price, we can use the market price as a cost measure. The total cost of the production factor in the cost-benefit analysis is, then, the market price multiplied by the number of units used. However, if the policy will also affect the market equilibrium, we cannot simply use the market price in our analysis. Figure 27.1 shows the difference.

Initial market equilibrium is at price $P_0$ and quantity $Q_0$. Imagine that a new policy will lead to a decrease in the market price to $P_1$, which increases quantity consumed to $Q_1$. The total benefits of this policy include the increased consumer surplus for the original consumers (rectangular area $P_0P_1AC$) as well as the benefit of the new consumers (triangle area ABC). Hence, the total benefit may be represented by the area $P_0P_1AB$, which shows that when a policy has a nonmarginal effect on the market price, we cannot simply use the prepolicy market price in a cost-benefit analysis.

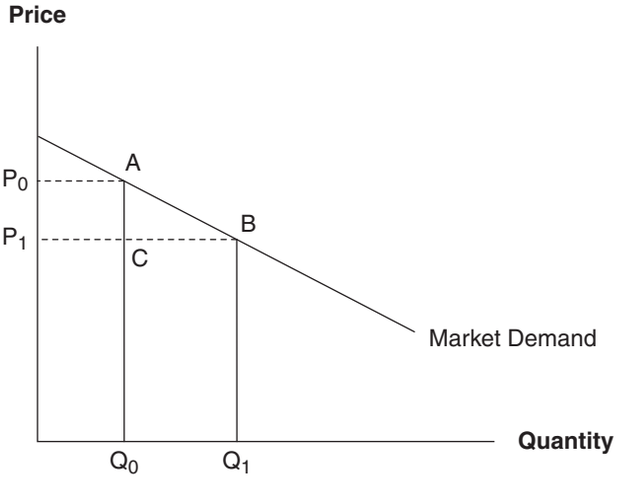


**Figure 27.1** Measuring Benefits and Costs for Nonmarginal Changes in Quantity

Other problems to note when using market data as information on prices in a cost-benefit analysis include market imperfections, such as tax distortions and externalities. The existence of taxes implies that there are several market prices, including or excluding taxes. An easy rule of thumb is to calculate prices including taxes, which implies that prices are expressed as consumer prices rather than producer prices. This implies that, for example, labor costs should be calculated as gross wages plus other social benefits or taxes that the consumer (or employer) has to pay. There is also another important effect of taxes that needs to be considered. Usually, cost-benefit analyses are performed for public projects, which often are paid for by taxes that lead to distortions in the economy. Distortions are created because not all activities are taxed, such as leisure. This implies that taxes on labor incomes change the relative prices between labor and leisure and lead to economic inefficiency. The distortion should normally be included in a cost-benefit analysis. As an example, in Swedish cost-benefit analysis, the recommendation is to include a distortion cost of 30% of direct costs (including taxes).

A final complicating note is that theoretically it may be that the correct benefit measure is the option price of a policy, which consists of the expected surplus (as outlined above) as well as the option value of a policy. Option value is the value that individuals are willing to pay, above the expected value of actually consuming the good, to have the option of consuming the good at some point in time (Weisbrod, 1964). For example, investing in a new national park includes the expected surplus of actually going to the park for the visitors. But it may also include a willingness to pay reflecting that the national park provides an option to go there, even if an individual will never actually go. In practical applications, it is common to exclude option value in the benefit calculations. There are several arguments for this; for example, it is difficult to measure and separate option values from other types of

values or attitudes that may not be relevant, such as nonuse values (non-use value refers to value an individual may place on, e.g., saving the rain forest in a specific region in South America, even if the individual knows that he or she will never go there—an intrinsic value). In addition to option value, sometimes it is also argued that a quasi-option value (sometimes referred to as *real option*) should be included in a project evaluation. The quasi-option value refers to the willingness to pay to avoid an irreversible commitment to the project right now, given expectations of future growth in knowledge relevant to the consequences of the project. It is not particularly common to include quasi-option value in cost-benefit analyses.

## Valuation of Nonmarket Goods

A common difficulty with a cost-benefit analysis is the fact that the policy to be evaluated includes a nonmarket good that is publicly provided. This implies that there are no market data to use. Examples of nonmarket goods that are common in cost-benefit analyses of public policy are values of safety, time, environmental goods, and pollution. Generally, there are two main approaches available for estimating monetary values of nonmarket goods: (1) revealed preference (RP) methods and (2) stated preference (SP) methods. RP methods use actual behavior and try to estimate implicit values of WTP or WTA. SP methods use surveys and experiments where individuals are asked to make hypothetical choices between different policy alternatives. Based on these choices, the researcher can estimate WTP or WTA.

### *Revealed Preference Method*

Revealed preference techniques can be used to elicit willingness to pay when there is market information about behavior that at least indirectly includes the good that the analyst is interested in evaluating.

One revealed preference approach is hedonic pricing (Rosen, 1974). Imagine that we would like to estimate the willingness to pay to avoid traffic noise; we may be able look at the housing market in a city to accomplish this. The price of a house depends on many different characteristics: size, neighborhood, number of bedrooms and bathrooms, construction year, and so on. But another important determinant may be the level of noise—that is, a house located close to a heavily trafficked highway will generally be less expensive than an identical house located in a noise-free environment. The hedonic pricing approach uses this intuition and performs a regression analysis where the outcome is the market value of a house including several relevant characteristics as determinants of the house price (including the noise level, measured in decibels). The results from such a statistical analysis can in a second step tell us the impact of the noise level on the market price, holding other important factors constant. For example, Nils Soguel (1994) uses data on monthly rent for housing in the city of Neuchâtel in Switzerland and included factors measuring the structure and condition of the building, several apartment-specific factors, and the location of the property. Based on a hedonic pricing approach, it shows that a one-unit increase in decibel led to a reduction in rents by 0.91%. Hence, using this approach, it is possible to estimate the economic value of noise.

Another revealed preference approach is the travel-cost method (see, e.g., Cicchetti, Freeman, Haveman, & Knetsch, 1971). This method uses the fact that individuals can reveal the value of a good by the amount of time they are willing to devote to its consumption. For example, if an individual pays $10 for a train ride that takes 30 minutes to visit a national park, we can use this information to indirectly estimate the lower bound of the value that the individual assigns the park. If the value of time for the individual is $20/hour, the individual is at least willing to spend the train fare of $10 plus $20 for the pure time cost (60 minutes back and forth) to visit the park—that is, a total sum of $30. Using this approach on a large set of individuals (or based on average data from different cities or regions), it is possible to estimate the total consumer surplus associated with the national park.

### *Stated Preference Method*

In many cases, it may not be possible to use revealed preference methods. Further, it should be noted that revealed preference methods assume that individuals (on average) have reasonable knowledge of different product characteristics for a hedonic pricing approach to give reliable estimates. In many cases, this condition may not be fulfilled. A possible option is then to turn to stated preference methods, which are based on hypothetical questions designed such that individuals should reveal the value they would assign to a good if it were implemented in the real world (Bateman et al., 2004). There are two common approaches in the stated preference literature: (1) contingent valuation (CV) and (2) choice modeling (CM). A CV survey describes a scenario to the respondent—for example, a proposed policy of investing in a new railway line—and asks the individual about his or her willingness to pay. A common recommendation is to use a single dichotomous-choice question—that is, respondents are asked whether they would be willing to pay $X for a project—and use a coercive payment mechanism (e.g., a tax raise) for the new public good (Carson & Groves, 2007). The cost of the project is varied in different subsamples of the study, which makes it possible to estimate the willingness to pay (demand curve) for the project using econometric analysis.

In a CM framework, a single respondent is asked to choose between different alternatives where different characteristics of a specific good are altered. For example, the respondent may choose between Project A, Project B, and status quo. Project A and Project B may be two different

railway line investments that differ with respect to commute time, safety, environmental pollution, and cost. Using econometric techniques, it is possible to estimate the marginal willingness to pay for all these different attributes using the choices made by the respondents.

Stated preference methods have the advantage that it is possible to directly value all types of nonmarket goods, but the reliability of willingness to pay estimates is also questioned by many economists. Problems include individuals' tendency to overestimate their willingness to pay in a hypothetical scenario compared to a real market scenario, referred to as *hypothetical bias* (Harrison & Rutström, 2008). There are also many studies that highlight the problem of *scope bias* (Fischhoff & Frederick, 1998), which refers to the fact that willingness to pay is often insensitive to the amount of goods being valued—for example, willingness to pay is the same for saving one whale as for saving one whale and one panda.

## Application: The Value of a Statistical Life

Many cost-benefit analyses concern public policies with effects on health risks (mortality and morbidity risks). Environmental regulation and infrastructure investment are two examples where policies often have direct impacts on mortality risks, morbidity risks, or both. Hence, we need some approach to monetize health risks. In the United States, an illustrative example can be found in the evaluation of the American Clean Air Act by the Environmental Protection Agency, where 80% of the benefits were made up of the value of reduced mortality risks (Krupnick et al., 2002). In a European example (from Sweden), cost-benefit analyses of road investments show that approximately 50% of the benefits consist of mortality and morbidity risk reductions (Persson & Lindqvist, 2003).

In the literature of the last 20 to 30 years, the concept used to monetize the benefit of reduced mortality risk is the value of a statistical life (VSL). It may be described in the following way:

> Suppose that you were faced with a 1/10,000 risk of death. This is a one-time-only risk that will not be repeated. The death is immediate and painless. The magnitude of this probability is comparable to the annual occupational fatality risk facing a typical American worker and about half the annual risk of being killed in a motor vehicle accident. If you faced such a risk, how much would you pay to eliminate it? (Viscusi, 1998, p. 45)

Let us assume that a certain individual is willing to pay \$100 to eliminate this risk. Using this information on WTP, the value of a statistical life is then based on the concept of adding up this total willingness to pay for a risk reduction of 1 in 10,000 to 1. Hence, in this example, it implies that the estimate for the value of a statistical life is equal to $\$100 \times 10{,}000 = \$1$ million (VSL = WTP / $\Delta$ risk). This implies that a policy that prevents one premature fatality increases social welfare as long as the cost is less than \$1 million.

What value should be used in a cost-benefit analysis? There is no market where we explicitly trade with small changes in mortality risks. Rather, researchers have been forced to turn to RP and SP methods to estimate VSL. The most common RP approach has been to use labor market data to estimate the wage premium demanded for accepting a riskier job (hedonic pricing). The idea behind these studies is that a more dangerous job will have to be more attractive in other dimensions to attract competent workers, and one such dimension is higher pay. Hence, by controlling for other important determinants of the wage, it is possible to separate the effect that is due to a higher on-the-job fatality risk. This approach has been particularly popular in the United States; see W. Kip Viscusi and Joseph Aldy (2003) for a survey of several papers using this approach.

SP approaches to estimate VSL are also frequent. Primarily, they have been performed using the contingent valuation approach. For example, a survey might begin with a description of the current state of the world regarding traffic accidents in a certain municipality, region, or country. The respondent might be told that in a population of 100,000 individuals, on average, 5 people will die in a traffic accident the next year. After this description, the respondent might be asked to consider a road safety investment that would, on average, reduce this mortality risk from 5 in 100,000 to 4 in 100,000—that is, 1 fewer individual killed per 100,000 individuals. To elicit the preferences of the respondent, the following question may be asked: Would you be willing to pay \$500 in a tax raise to have this traffic safety program implemented? The respondent then ticks a box indicating yes or no. Other respondents are given other costs of the project, which gives the researcher the possibility of estimating a demand curve for the mortality risk reduction.

In the United States, the Environmental Protection Agency recommends a VSL estimate of €6.9 million for cost-benefit analyses in the environmental sector. In Europe, for the Clean Air for Europe (CAFE) program, a VSL (mean) of €2 million (approx. \$2.7 million) is suggested (Hurley et al., 2005). Theoretically, higher income and higher baseline risk should be associated with a higher VSL (although this can hardly explain the large differences between the estimates used by the United States and the European Union). In the transport sector, there are also international differences. Among European countries, Norway recommends a VSL estimate for infrastructure investments of approximately €2.9 million; the United Kingdom, €1.8 million; Germany, €1.6 million; Italy, €1.4 million; and Spain, €1.1 million (HEATCO, 2006).

For many individuals, it is offensive to suggest that the value of life should be assigned a monetary value, and there is some critique from the research community (Broome, 1978). However, it needs to be acknowledged

that in estimating WTP for small mortality risk reductions for each individual (hence the term *statistical life*), no one is trying to value the life of an identified individual. Moreover, these decisions have to be made, and we make them daily on an individual basis. Public policy decision making on topics that have impacts on mortality risks will always implicitly value a prevented fatality. Using the concept VSL in cost-benefit analysis, economists are merely trying to make these decisions explicit and base them on a rational decision principle.

## Discounting

As stated in the introduction, benefits and costs associated with a policy that occur at different times need to be expressed in a common metric. This metric is the present value of benefits and the present value of costs. Practically, discounting into present value is calculated as $PV = B_t / (1 + SDR)^t$, where $PV$ is the present value of a benefit ($B_t$) occurring in year $t$ in the future. $SDR$ is the social discount rate. For example, with a social discount rate of 3%, a benefit of \$100 occurring in 5 years has a present value of $\$100 / (1 + 0.03)^5 = \$86.26$. Traditionally, there have been two main approaches to choosing an appropriate social discount rate: (1) the social opportunity rate cost of capital, and (2) social time preference rate. The former can be seen as the opportunity cost of capital used for a certain policy. Imagine that a road safety investment has a cost of \$100; this money could instead be placed in a (more or less) risk-free government bond at a real interest rate of perhaps 5%. This would be the opportunity cost of the capital used for the public policy. The social time preference rate approach can be formulated as in the optimal growth model (Ramsey, 1928) outlining the long-run equilibrium return of capital: $SDR = \rho + \mu \times g$, where $\rho$ is the pure time preference of individuals, $\mu$ is the income elasticity, and $g$ is the growth rate of the economy. This captures both that individuals tend to receive \$1 today rather than in a year ($\rho$), and it reflects that as individuals grow richer, each additional \$1 is worth less to them ($\mu$).

The actual discount rate used in economic evaluation often has a major impact on the result. Consider Table 27.1, which shows the present value of \$1 million in 10, 30, 50, and 100 years in the future with discount rates of 1%, 3%, 5%, and 10%. As an example, with a discount rate of 1%, the present value of \$1,000,000 occurring in 10 years is \$905,287. The present value if using a discount rate of 5% is instead \$613,913.

Table 27.1 can be used to show how important the discount rate is for the discussion regarding what to do about global warming. The predicted costs of global warming are assumed to lie quite distantly in the future. The effect of different discount rates will be relatively larger the more distant in the future the benefit or cost will take place. An environmental cost of \$1 million occurring in 100 years is equal to only \$7,604 in present value if using a discount rate of 5%. If using a discount rate of 1%, it is \$369,711. A policy that would be paid for today to eliminate this cost in 100 years would be increasing social welfare if it cost less than \$7,604 using the higher discount rate (5%) or would be increasing social welfare if it cost less than \$369,711 using the lower discount rate (1%). Hence, it is obvious that the conclusion on how much of current GDP we should spend to decrease costs of global warming occurring in the distant future will be highly dependent on the chosen discount rate used in the cost-benefit analysis.

In 2006, a comprehensive study on the economics of climate change was presented by the British government: the Stern Review (Stern, 2007). The review argues that the appropriate discount rate for climate policy is 1.4%. Stern argues that because global warming is an issue affecting many generations, the pure time preference ($\rho$) should be very low (0.1), and he assumes an income elasticity ($\mu$) of 1 and a growth rate ($g$) of 1.3; this implies $SDR = 0.1 + 1 \times 1.3 = 1.4$. This is a relatively low discount rate compared to what most governments recommend for standard cost-benefit analyses around the world for projects with shorter life spans than policies to combat global warming. The Stern Review has also been criticized by other economists, who argue that such a low discount rate is not ethically defensible and has no connection to market data or behavior (Nordhaus, 2007). Weitzman (2007) argues that a more appropriate assumption is that $\rho = 2$, $\mu = 2$, and $g = 2$, which would give a social discount rate of 6%. The debate about the correct social discount rate has been a very public question in discussions about global warming policy—that is, how much, how fast, and how costly should measures taken today to reduce carbon emissions be?

**Table 27.1** Present Value of \$1 Million Under Different Time Horizons and Discount Rates

| *Discount Percentage Rate* | *10 Years* | *30 Years* | *50 Years* | *100 Years* |
|---|---|---|---|---|
| 1 | \$905,287 | \$741,922 | \$608,038 | \$369,711 |
| 3 | \$744,093 | \$411,987 | \$228,107 | \$52,032 |
| 5 | \$613,913 | \$231,377 | \$87,204 | \$7,604 |
| 10 | \$385,543 | \$57,309 | \$8,519 | \$73 |

There really exists no consensus regarding the correct discount rate, and there probably never will. Different government authorities around the world propose different social discount rates. The European Union demands that cost-benefit analyses be conducted for projects that imply important budget consumption, and the European Commission proposes a social discount rate of 5%. In the so-called Green Book in the United Kingdom, a social discount rate of 3.5% is proposed. In France, the Commisariat Général du Plan proposes a discount rate of 4%. In the United States, there are somewhat different proposals for different sectors, but the Office of Management and Budget recommends a social discount rate of 7% (Rambaud & Torrecillas, 2006). Several of the recommendations are also indicating that the social discount rate should be dependent on the time horizon of the project; in the United Kingdom, the social discount rate is proposed to be 3.5% for year 0 to 30, 3% for year 31 to 75, and decreasing down to 1% for policies with a life span of more than 301 years.

## Sensitivity Analysis

There are often large uncertainties in a cost-benefit analysis, regarding parameter estimates of benefits and costs. It has been shown that, especially for large projects, costs are often underestimated and benefits are sometimes exaggerated, making projects look more beneficial than they actually are (Flyvbjerg, Holm, & Buhl, 2002, 2005). These types of uncertainties need to be explicitly discussed and evaluated in the analysis. One common approach to deal with uncertainty in a cost-benefit analysis is to perform sensitivity analyses. There are different approaches regarding how to conduct a sensitivity analysis. Partial sensitivity analysis involves changing different parameter estimates and examines how it affects the net present value of the policy. Examples include using different discount rates and different parameter values of the value of a statistical life. Another approach is the so-called worst- and best-case analysis. Imagine that the benefits are uncertain but that an interval can be roughly estimated—for example, the benefit of improving environmental quality will be in the interval $100,000 to $150,000. A worst-case analysis implies taking the lowest bound of all beneficial parameter estimates. A best-case analysis implies the opposite. These types of sensitivity analyses may be interesting for a risk-averse decision maker and also give information about the lowest benefit (or largest loss) for a given project.

The downside to partial sensitivity analysis and worst- and best-case scenarios is that they do not take all available information about the parameters into consideration. Further, they do not give any information about the variance of the net present value of a project. For example, if two projects give similar net present value, decision makers may be more interested in the project with the lowest variance around the outcome. This requires the use of Monte Carlo sensitivity analysis. This is based on simulations where economists make assumptions about the statistical distribution of different parameters and perform repeated draws of different parameter values, each leading to a different net present value. This can give an overview of the distribution of the uncertainty of the project. A standard approach to visually describe the results from a Monte Carlo sensitivity analysis is to display the results in a histogram that shows mean net present value, sample variance, and standard error.

## An Application

To end this overview, a cost-benefit analysis of the Stockholm congestion charging policy is described (Eliasson, 2009). The Stockholm road congestion charging system is based on a cordon around central parts of Stockholm (capital of Sweden), with a road toll between 6:30 a.m. and 6:30 p.m. weekdays (higher charge during peak hours). The aim with the charging system is to reduce congestion and increase the reliability of travel times. Positive effects on safety and the environment are also expected. The cost-benefit analysis has the particular advantage of being based on observed traffic behavior, rather than simulations and forecasts (this is possible because the charging system was introduced during a trial period of 6 months). Using the six steps in a cost-benefit analysis as outlined in the introduction:

1. The first step involves defining the relevant population. The Stockholm congestion charging policy is mainly relevant for the population in the region of Stockholm, but because the cost-benefit analysis is based on actual behavior and data, it will therefore include benefits and costs of users of the roads in Stockholm, which will include various types of visitors as well. Hence, the relevant population is all the users of the roads.
2. The second step in a cost-benefit analysis is to identify the relevant consequences associated with the policy. The following main benefits are associated with the policy: (a) reduction in travel times due to decreased congestion, (b) increased reliability in travel times, (c) reductions of carbon dioxide and health-related emissions due to the decrease in traffic volume, and (d) increased road safety due to decrease in traffic volume. It could be hypothesized that the system would have effects on decisions where to locate, the regional economy, and retail sales, but it has been judged that these effects will be very small. Negative effects are (a) investment and startup costs, and (b) yearly operation costs.
3. The third step in the cost-benefit analysis is to monetize all the benefits and costs associated with the policy. Table 27.2 summarizes the consequences and shows their monetary benefits and costs. Some of the smaller benefits and costs in the analysis are not described here; refer to the reference for a more detailed description.

**Table 27.2** Annual Benefits and Costs of the Stockholm Road Charging System

| | *Consequence* | *Value per Unit* | *Total Benefit or Cost* |
|---|---|---|---|
| **Benefits** | | | |
| Decrease in travel time | Approximately 4.4 million hours | 122 SEK/hour | 536 million SEK |
| Increased reliability in travel time | — | 0.9 × (value of time) × (standard deviation of travel time) | 78 million SEK |
| Decrease in emissions | Decrease in $CO_2$ by 2.7% | 1.50SEK/kg $CO_2$ | 64 million SEK (+ 22 million SEK for health-related effects) |
| Decrease in traffic fatalities and injuries | Decrease by 3.6% | 17 million SEK per fatality | 125 million SEK |
| **Costs** | | | |
| Operation costs | Reinvestment and maintenance | Machine hour costs and labor hour costs | 220 million SEK |
| Marginal cost of public funds | Distortionary tax effects | 1.3 × all direct costs | 182 million SEK |
| **Net social benefit excluding investment costs** | | | **654 million SEK (91 million USD)[a]** |

SOURCE: Based partially on data reported in Eliasson (2009).

NOTE: a. Based on the December 16, 2009, exchange rate: 1 USD = 7.2 SEK.

Table 27.2 shows the annual benefits and costs of the system. The magnitude of the effects on travel time, standard deviation of travel time, and so on is based on large computer estimations of the traffic measurements on 189 links to calibrate origin-destination (OD) matrices for the case with and without the charging system. The investment cost is not listed in Table 27.2 but was estimated at 1.9 million SEK.

4. The next step consists of calculating the net present value of all benefits and costs based on the annual estimates. It is not obvious which time horizon should be used for the project, but technical data and past experience indicated that it was reasonable to make a conservative assumption that the system would have an economic life span of 20 years. Hence, the benefits and operating costs in year 2, 3, . . . , 20 have to be discounted to a present value. The Swedish National Road Administration argues that the social discount rate should be 4%. Hence, the present value of the net social benefits in year 20 is $654 / (1.04)^{20} = 298$ million SEK. These calculations are performed for benefits and costs in years 1 through 20.
5. Discount all annual benefits and costs to present values, as in Step 4, showing that the total social surplus (after deducting the investment costs) is approximately 6.3 billion SEK (approx. $800 million). Expressing it as a payback estimate, this means that the policy will take 4 years before the investment costs are fully repaid. It is also explicitly discussed that some consequences were deemed too difficult to include in the calculations, such as the effects on noise, labor market, time costs for users, quicker bus journeys, and so on.
6. The sixth step in a cost-benefit analysis is to perform a sensitivity analysis. In this aspect, there is little done in the described analysis. One reason for this is that the actual estimates of the consequences as performed using OD matrices are very time consuming, which more or less implies that because of practical limitations, only one main estimation can be performed. A simple sensitivity analysis is performed assuming increasing benefits over the time horizon. But if any improvement to the cost-benefit analysis should be suggested, it would be to conduct a more detailed sensitivity analysis. The quite straightforward conclusion of the cost-benefit analysis, even though it should have included sensitivity analyses to satisfy our full requirements, is that social welfare will increase because of the charging system.

## Conclusion

How should we evaluate a proposed public policy or regulation? A cost-benefit analysis is an approach that includes all relevant consequences of a policy and compares, in monetary units, benefits with costs. If benefits outweigh the costs, the policy is said to increase social welfare. Social welfare is defined using the Hicks-Kaldor criterion, which states that a policy increases welfare if the winners from the policy can compensate the losers from the policy and still be better off than if the policy is not implemented.

To conduct a cost-benefit analysis, one must identify consequences and express them in a monetary metric so that all consequences can be compared in the same unit of measurement. If benefits and costs arise in the future, they should be discounted to present value using a social discount rate. Finally, given the uncertainties involved with estimating consequences of a policy or regulation as well as uncertainties with the monetary estimates of the consequences, a proper cost-benefit analysis should include sensitivity analyses to show how robust the result is.

Finally, considering the definition of social welfare as usually applied in cost-benefit analysis (Hicks-Kaldor criterion), the typical cost-benefit analysis of a project or regulation estimates the effect on economic efficiency. Therefore, even though a very important guide to decision making, in most applications, cost-benefit analysis is often seen as one of several guides to the decision making process. Especially in political decision making, there will be other effects of interest, such as effects on income distribution and geographical distribution of benefits.

## References and Further Readings


Bateman, I. J., Carson, R. T., Day, B., Hanemann, N., Hett, T., Hanley, N., et al. (2004). *Economic valuation with stated preference techniques: A manual.* Cheltenham, UK: Edward Elgar.

Broome, J. (1978). Trying to save a life. *Journal of Public Economics, 9,* 91–100.

Carson, R. T., & Groves, T. (2007). Incentive and informational properties of preference questions. *Environmental and Resource Economics, 37,* 181–210.

Cicchetti, C. J., Freeman, A. M., Haveman, R. H., & Knetsch, J. L. (1971). On the economics of mass demonstrations: A case study of the November 1969 march on Washington. *American Economic Review, 61,* 719–724.

Eliasson, J. (2009). A cost-benefit analysis of the Stockholm congestion charging system. *Transportation Research Part A, 43,* 46–480.

Fischhoff, B., & Frederick, S. (1998). Scope (in)sensitivity in elicited valuations. *Risk, Decision, and Policy, 3,* 109–123.

Flyvbjerg, B., Holm, M. K., & Buhl, S. L. (2002). Underestimating costs in public works projects: Error or lie? *Journal of the American Planning Association, 68,* 279–295.

Flyvbjerg, B., Holm, M. K., & Buhl, S. L. (2005). How (in)accurate are demand forecasts in public works projects? The case of transportation. *Journal of the American Planning Association, 71,* 131–146.

Harrison, G. W., & Rutström, E. E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods. In C. Plott & V. Smith (Eds.), *Handbook of experimental economics results* (pp. 752–767). New York: Elsevier Science.

HEATCO. (2006). Deliverable 5 proposal for harmonised guidelines. *Developing harmonised European approaches for transport costing and project assessment* (Contract No. FP6–2002-SSP-1/502481). Stuttgart, Germany: IER.

Hurley, F., Cowie, H., Hunt, A., Holland, M., Miller, B., Pye, S., et al. (2005). Methodology for the cost-benefit analysis for CAFE: Volume 2: Health impact assessment. Available at http://www.cafe-cba.org/assets/volume_2_methodology_overview_02-05.pdf

Kahneman, D., Knetsch, J. L., & Thaler, R. (1990). Experimental tests of the endowment effect and the Coase Theorem. *Journal of Political Economy, 98,* 1325–1348.

Krupnick, A., Alberini, A., Cropper, M., Simon, N., O'Brien, B., Goeree, R., et al. (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,* 161–186.

Nordhaus, W. (2007). A review of the *Stern Review on the economics of climate change. Journal of Economic Literature, 45,* 686–702.

Persson, S., & Lindqvist, E. (2003). *Värdering av tid, olyckor och miljö vid vägtrafikinvesteringar—Kartläggning och modellbeskrivning* [Valuation of time, accidents and environment for road investments: Overview and model descriptions] (Rapport 5270). Stockholm, Sweden: Naturvårdsverket (Swedish Environmental Protection Agency).

Rambaud, S. C., & Torrecillas, M. J. (2006). Social discount rate: A revision. *Anales de Estudios Económicos y Empresariales, 16,* 75–98.

Ramsey, F. (1928). A mathematical theory of saving. *Economic Journal, 38,* 543–559.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82,* 34–55.

Soguel, N. (1994). *Evaluation monetaire des atteinties a l'environment: Une etude hedoniste et contingente sur l'impact des transports* [Economic valuation of environmental damage: A hedonic study on impacts from transport]. Neuchatel, Switzerland: Imprimerie de L'evolve SA.

Stern, N. (2007). *The economics of climate change.* Cambridge, UK: Cambridge University Press.

Weisbrod, B. A. (1964). Collective-consumption services of individual-consumption goods. *Quarterly Journal of Economics, 78,* 471–477.

Weitzman, M. (2007). A review of the *Stern Review on the economics of climate change. Journal of Economic Literature, 45,* 703–724.

Willig, R. D. (1976). Consumer surplus without apology. *American Economic Review, 66,* 589–597.

Viscusi, W. K. (1998). *Rational risk policy.* Oxford, UK: Oxford University Press.

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,* 5–76.

# PART IV

## Macroeconomics

# 28

# Economic Measurement and Forecasting

Mehdi Mostaghimi

*Southern Connecticut State University*

In December 2000, within a few days after a U.S. Supreme Court decision on the result of the November presidential election, then–Federal Reserve Chairman Alan Greenspan announced a grim prediction of a downturn for the near future of the U.S. economy. To counter it, the Federal Reserve implemented a series of expansionary monetary policy actions—cutting the federal funds rate and increasing the supply of money for the ensuing 2 years. In July 2003, the National Bureau of Economic Research (NBER), a nongovernmental think tank organization consisting mostly of academicians, identified and dated a recession in 2001 for the United States with a starting date of March and an ending date of November, for a total of 7 months.

An interested person learning about macroeconomic policy decision making, after reading the preceding paragraph, curiously asks a series of questions focused on identifying economic activities, measuring economic activities, modeling information, economic forecasting, and decision and policy making. The focus of this chapter is to briefly provide answers to some of these questions without getting into the complexities inherent in their details.

A *recession* is a contraction or downturn in overall economic activity when output and employment decline, generally for a period of 6 months (two quarters) or more. A severe and long contraction is called a *depression.* A relatively mild and short contraction (less than 6 months) is referred to as a *downturn* or *slowdown.* When economic activities change at a positive rate (output and employment are rising), the economy is said to be in an *expansion* phase. When the demand for the goods and services in an economy exceeds its ability to meet them or when for some other reason prices throughout the economy rise, the economy experiences *inflation. Stagflation* occurs when an economy experiences high unemployment and inflation simultaneously.

The preceding phases of the economy reflect the business cycle (Figure 28.1). Each economy has a long-run growth rate that is dependent on its economic structure and a range of socioeconomic and technological factors. In a dynamic economy, as these factors change—for example, through technological advancements—the long-run growth rate will change. In the short run, economic growth may be rising or falling around the long-run growth rate. When the short-run growth rate is falling, the economy is in a slowdown that may lead to a recession. When the short-run growth rate is rising, the economy is in an expansion. These are the phases of the business cycle. A *business cycle* consists of an expansion and a contraction that follow each other separated by a peak and a trough. A *peak,* the high point of a cycle, is the beginning of a downturn, and a *trough,* the low point of a cycle, is the beginning of an expansion. An expansion may contain one or more downturns and slowdowns. Not all the slowdowns and downturns end up in a recession, and predicting if and when a recession will occur is difficult.

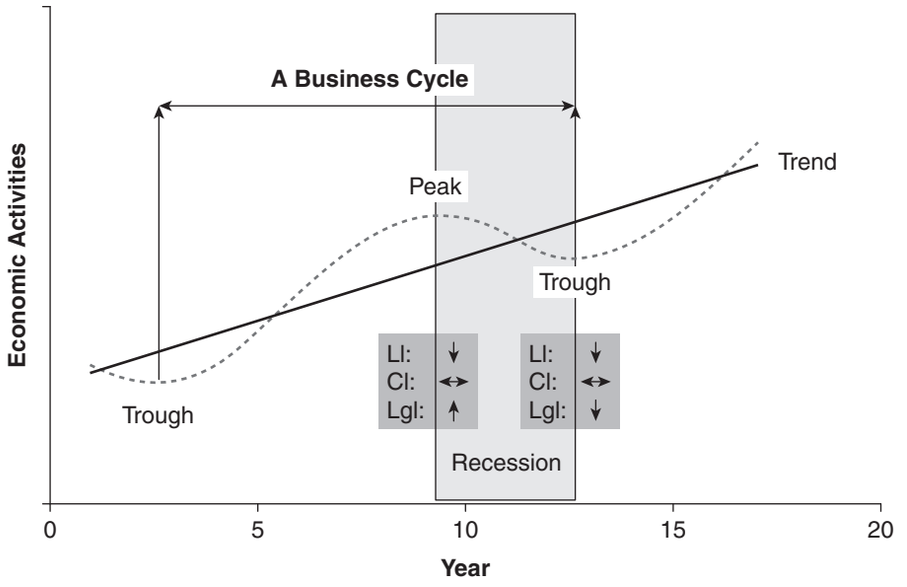


**Figure 28.1** The Business Cycle

NOTES: A *business cycle* consists of an expansion and a contraction that follow each other, separated by a peak and a trough. A *peak* is identified when the changes in leading indicators are negative, no changes are observed in the coincidental indicators, and the changes in the lagging indicators are positive. A *trough* is identified when the changes in the leading indicators are positive, no changes are observed in the coincidental indicators, and the changes in the lagging indicators are negative.

From 1959 to 2007, the U.S. economy has gone through seven business cycles with a cycle averaging 7 years, expansions averaging 6 years, and recessions averaging 1 year. From 1982 to 2007, there have been two complete business cycles ending in March 1991 and November 2001 with an average of 10 years and with expansion and recession averages of 9 years and 8 months, respectively. In this period, there were several downturns that did not turn into recessions, notably in 1987 and 1996. Longer expansion periods and shorter recession periods have been characteristics of the modern U.S. economy, reflecting the impact of advancements in predicting peaks and troughs accurately and of implementing the appropriate monetary and fiscal policies in a timely manner.

For economic policy decision making (monetary and fiscal), predicting a turning point (peak and trough) in a business cycle is very important. A main objective of economic policy is to create a stable economy that grows at a desirable rate with high employment and low inflation. Policy actions designed to bring stability in the economy are business cycle dependent. For example, to prevent an economy from overheating during an expansion period, which may result in rapid price hikes for goods and services and a high inflation rate, policy actions such as tightening the money supply and raising interest rates will cut the demand for goods and capital and cool the economy down. An example of policy action for preventing a slowing economy from entering a recession is a combination of fiscal stimulus, such as a tax cut, and easy money, such as lower interest rates, to encourage customer and business spending.

Predicting a downturn or the peak that it will follow is considered critically important for policy making because a peak is the start of a recession. If an impending recession can be predicted correctly and appropriate policy actions are designed and implemented in a timely manner, a recession and the resulting unemployment and business failures can be entirely avoided, or their severity and duration reduced.

Expansionary monetary policy actions implemented by the Federal Reserve after correctly predicting an impending recession in December 2000 did not prevent it from happening in 2001, but those actions likely reduced the potential severity and duration of that recession and maintained economic stability despite the economic shock of the September 2001 World Trade Center attack.

To formulate an effective economic policy for a country, it is critical to correctly measure the economy's performance. The parameters used to gauge the performance of the economy and their measurements are generally called *economic indicators* and *economic indexes*.

An economic system can be monitored using parameters that describe its different aspects, such as income, production, employment, inflation, and so on. For each parameter, there may be several indictors that are used to measure it. For example, both the consumer price index and producer price index are measures of inflation. As in any dynamic system, not all of the economic parameters are known a priori; they are discovered and their measurements are invented as the economy evolves.

By modeling and analyzing these indicators, it is possible to study the behavior of an entire economic system. Forecasting the future behavior of the state of an economy

in a business cycle is a projection of its past behavior onto the future. Because the future will inevitably vary from the past, no forecast is perfectly accurate, and uncertainty is associated with every forecast.

The focus of this chapter is on economic indicators and their use in predicting a downturn (or the peak that it will follow) in the United States. The next section focuses on describing and grouping economic indicators in the United States. This is followed by a section on modeling to combine the information of these indicators for predicting a turning point in the overall state of the economy. Discussions focused on forecast performance and evaluation and on economic policy decision making are also presented in this section. The final section focuses on future directions of economic indicators and on forecasting in the twenty-first century.

## Economic Indicators

An economic *parameter* describes an aspect or dimension of an economic system. An economic *indicator* is a measure of a parameter for an economy at a given point in time. A parameter may have several indicators, each measuring it differently. All the parameters of an economy together describe the overall behavior of that system. A collection of all economic indicators describes the behavior of an entire economic system at a given point in time (Burns & Mitchell, 1946).

Gross domestic product (GDP), which measures the total value of all goods and services produced in an economy, is probably the most important indicator of the performance of an economy. U.S. GDP for the first quarter of 2008 was estimated at $11.6 trillion. The change in the value of GDP from one quarter to another, in percentage form, is a measure of economic growth. When GDP is adjusted (deflated) for price changes, it is called *real GDP*. A positive percentage change in real GDP is an indication of expansion, and a negative percentage change is an indication of a downturn, or contraction. Real GDP grew by 0.9% in the first quarter of 2008, indicating a weak expansion. To study the behavior of an economy, the change in the value of an indicator is more informative than the indicator's overall value.

A few other major parameters and their main indicators for the U.S. economy are (a) inflation and price stability, measured by consumer price index and producer price index; (b) employment and labor force use, measured by the unemployment rate; (c) efficient use of the employed labor force, measured by the productivity index; (d) cost of borrowing money, measured by the prime interest rate; and (e) consumers' future economic behavior, measured by the consumer sentiment index.

In the United States, economic indicators are mostly produced by such federal government agencies as the Commerce Department, the Bureau of Labor Statistics, and the Federal Reserve Bank. The U.S. Congress compiles and publishes monthly reports on most of these indicators. A few indicators are also produced by major universities, such as the University of Michigan, and by economic research organizations, such as the Conference Board.

Each state in the United States has similar economic parameters and indicators focused on its towns, cities, and counties. State government agencies, such as state departments of labor and economic development, as well as state universities, economic research organizations, and chambers of commerce, are developers and producers of economic indicators. In the last 10 to 15 years, major advancements have been made in this area at the state and local levels.

There are many ways to group economic indicators for systematically studying an economy. Three of them are presented here:

1. Grouping by *economic category*. The broad categories of economic activity are the base for this grouping, which includes categories such as income and output, employment and wages, and producer and consumer price.
2. Grouping by *information source*. This is an indicator of an aspect of the real (production) economy, money economy, or behavioral economy.
   - Real economy indicators are measures of the actual products and services produced and of economic factors such as labor and capital. GDP and the unemployment rate are two examples of real economy indicators. Some of these measures, such as GDP, are adjusted, or deflated, for price changes and inflation.
   - Money economy indicators are measures of the availability of money in the economy and of the security market's valuations. Measures of the money supply, the Federal Funds Rate, Treasury Bill rates, and the Standard & Poor 500 Index (S&P500) are examples.
   - Behavioral economy indicators are measures of consumers' and businesses' attitudes toward the future directions of the state of the economy in terms of being optimistic, pessimistic, or in between. These attitudes will affect consumers' and businesses' decisions about purchasing goods and services and investing in new equipment and machinery. The University of Michigan's Consumer Sentiment Index and the Conference Board's Consumer Confidence Index and CEO Confidence Survey are three examples here.
3. Grouping by *business cycle timing*. There are many indicators whose values are informative in identifying the location of an economy on the business cycle. They are lagging indicators, coincidental indicators, and leading indicators.
   - *Lagging indicators* identify the location of the economy on the business cycle in the near past. The peak and trough of a lagging indicator are the confirmations of the peak and trough of the recent state of an economy. For example, when a lagging indicator peaks, it confirms that the overall economy has already peaked. Average duration of unemployment is a lagging indicator. It peaked in April 2001, a month after the start of the 2001 recession.

- *Coincidental indicators* identify the current state of the economy in a business cycle. Real GDP is a coincidental indicator measuring the current, quarterly, value of goods and services produced in the economy. Real GDP peaked as early as the third quarter of 2000 to signal the start of the 2001 recession.
- *Leading indicators* are predictors of the near future state of the economy. Slope of the yield curve, defined as the interest rate spread between the 10-year treasury bonds and the federal funds, is a leading indicator. A negative slope of the yield curve is an indication of an impending slowdown and, possibly, a recession in the economy. It happens when the financial risks in the near future exceed the ones in the long term. This slope has been negative prior to all the recessions since 1959.

The list of indicators used to study the U.S. business cycle varies from one research organization to another. For example, the American Institute for Economic Research (AIER) uses different indicators than the Conference Board. The list of indicators used by the AIER is shown in Table 28.1, and those used by the Conference Board for leading indicators are shown in Table 28.2. These lists are, however, highly overlapped for each group and are modified occasionally once the relative performances of the indicators change. For example, the Conference Board removed 3 indicators from its list of leading indicators in 1996 and added 2 new ones to cut the total number of its leading indicators from 11 to 10.

A change in percentage in the value of an indicator is used for locating the economy on a business cycle. A change consists of a sign and a magnitude. For most of the indicators, a positive sign in the percentage change is an indication of a growth. For a few, such as the unemployment rate, a positive sign is an indication of a decrease in growth. The magnitude of a percentage change is a measure of the degree of impact. A large percentage increase in real GDP is a measure of strong economic growth, and a large percentage increase in the unemployment rate is a measure of a large contraction in an economy.

Identifying a peak and a trough in the economy is critical for predicting the start of a recession and the start of an expansion, respectively. Generally, a peak is identified when the changes in leading indicators are negative, no changes are observed in the coincidental indicators, and the changes in the lagging indicators are positive. A trough is identified when the changes in the leading indicators are positive, no changes are observed in the coincidental indicators, and the changes in the lagging indicators are negative.

Not all indicators of a group show the same sign and the same magnitude percentage change in a given month. Because it is possible for the indicators of a group to show conflicting signals on both the sign and the magnitude, a consensus of them is usually used for identifying the location of an economy on the business cycle.

For example, if five coincidental indicators are used to identify the current state of an economy, two extreme possibilities are for all to be either positive or negative with both having high percentage changes in magnitude. The interpretations of these two cases are straightforward in that the economy is strongly growing and severely contracting, respectively. However, if three of the five coincidental indicators show positive percentage changes, one has zero change, and one has a negative percentage change, and for the ones with nonzero percentage changes, the magnitude of the change varies from a small to a large value, a consensus interpretation of the values of these indicators for the state of the economy is not straightforward. Because, in addition to sign and magnitude of the change, other factors

**Table 28.1** Grouping Economic Indicators by Business Cycle Timing

The Primary Leading Indicators:

- M1 Money Supply
- Yield Curve Index
- ISM Price Index
- New Orders for Consumer Goods
- New Orders for Core Capital Goods
- New Housing Permits
- Ratio of Manufacturing and Trade Sales to Inventories
- Vendor Performance, Slower Deliveries Diffusion Index
- Index of Common Stock Prices
- Average Workweek in Manufacturing
- Initial Claims for State Unemployment Insurance
- 3-Month Percent Change in Consumer Debt

The Primary Roughly Coincident Indicators:

- Nonagricultural Employment
- Index of Industrial Production
- Personal Income less Transfer Payments
- Manufacturing and Trade Sales
- Civilian Employment as a Percentage of the Working-Age Population
- Gross Domestic Product

The Primary Lagging Indicators:

- Average Duration of Unemployment
- Manufacturing and Trade Inventories
- Commercial and Industrial Loans
- Ratio of Debt to Income
- Percent Change from a Year Earlier in Manufacturing Labor Cost per Unit of Output
- Composite of Short-Term Rates

SOURCE: AIER (2008). Reproduced with permission.

**Table 28.2** U.S. Composite Leading Indexes: Components and Standardization Factors: February 2008

| *Leading Indictors* | *Factor* |
|---|---|
| Average weekly hours, manufacturing | 0.2565 |
| Average weekly initial claims for unemployment insurance | 0.0310 |
| Manufacturers' new orders, consumer goods and materials | 0.0763 |
| Vendor performance, slower deliveries diffusion index | 0.0672 |
| Manufacturers' new orders, nondefense capital goods | 0.0186 |
| Building permits, new private housing units | 0.0270 |
| Stock prices, 500 common stocks | 0.0384 |
| Money supply, M2 | 0.3530 |
| Interest rate spread, 10-year treasury bonds less federal funds | 0.1037 |
| Index of consumer expectations | 0.0283 |
| Total | 1.0000 |

SOURCE: Reproduced with permission from the Conference Board Leading Economic Index™ (2008). © 2009 The Conference Board, Inc.

such as the relative importance of the indicators is also critical for coming up with an interpretation.

Economic research organizations use various schemes and models for generating a consensus interpretation for each group. For example, if the information used is limited to just the sign and magnitude of a change, the Conference Board produces diffusion indexes for all three groups. In addition, if the relative importance of the individual indicators is used, the Conference Board has composite indexes, such as composite leading economic indicators. AIER has similar indexes derived using different schemes. These factors are also used informally, usually by economic experts with extensive experience in economic forecasting, to produce a consensus interpretation for a group. Thus, given the high degree of subjectivity involved in this process, economists often arrive at different interpretations of the same information and produce different assessments of the state of the economy.

Another way to use these indicators to identify the position of an economy in a business cycle is to combine the information and interpretations of all three groups of indicators. If the changes in indicators of all three categories behave the same way, either positively or negatively, it is an indication that the economy is in a given state of the business cycle: expansion if the changes are positive and recession if the changes are negative. However, when the signs are different, they may point to a peak or trough of the cycle. In an expanding economy, negative changes in leading indicators, no change in coincidental indicators, and positive changes in lagging indicators are a sign of a possible peak in the economy. Similarly, in a declining economy, positive changes in the leading indicators, no change in the coincidental indicators, and negative changes in the lagging indicators are a sign of a trough in the economy. There are formal, econometric methods used for such analysis, which are briefly mentioned in the next section. However, it is the subjective interpretations of the information by the expert economists that are frequently used for decision making—a method that may result in even more conflicting interpretations!

## Combining Information and Predicting a Turning Point

In the previous section, economic indicators were described. We noticed that some of these indicators can be grouped as lagging indicators, coincidental indicators, and leading indicators. We have also noted that the values of these indicators could be used in a group to describe the overall state of an economy at a given time.

In this section, we focus on combining information from these indicators and using them to predict the near future state of an economy in a business cycle. The idea is that, for example, if there are 10 leading indicators that all individually describe the behavior of the state of an economy in the near future, a composite, a combination, of them should have a superior performance over the performance of each individual one—superior in yielding more accurate predictions and signaling fewer false alarms. A false alarm would be an indicator that signals a recession when in fact one is not likely to occur.

The question then becomes how to combine the information provided by the indicators and how to evaluate the performance of a composite indicator in identifying the state of an economy on a business cycle. Modeling and nonmodeling approaches are used in response to these questions. Both approaches are briefly described in this section. A detailed treatment of these approaches is quite mathematical, requiring a high level of understanding of mathematical and econometric modeling—something that is well beyond the scope of this introductory chapter (Lahiri & Moore, 1991). For interested readers who have a good technical grasp of economic modeling, a rather comprehensive treatment of these topics is the excellent collection of work presented in Graham Elliott, Clive Granger, and Allan Timmermann (2006).

### Modeling Overall Economic Performance

To model an economy's overall performance, several measures are used to capture different areas of economic activity. For example, because employment and production are two very important aspects of an economy, measures and indicators of these aspects, such as unemployment rate and industrial production, are both used in the analysis.

Because each measure provides information about a particular aspect of the economy, for studying an economy's overall behavior, a combination of these measures, a composite indicator, must be used.

GDP is the one indicator that is generally recognized as a good measure of the overall state of an economy. A disadvantage of GDP, however, is its low production frequency (quarterly as opposed to monthly). Other monthly indicators such as industrial production (IP), real personal income, and employment can, however, be used as proxies for GDP. These proxy indicators are not perfect, and their relative importance continually changes due to changes in the structure of an economy.

A more appropriate measure of the overall economic performance seems to be a composite measure of several indicators. The National Bureau of Economic Research (NBER) uses a combination of such measures as IP, employment, income, and sales in an informal, committee process for making decisions on business cycle dating. The Conference Board uses a formal, statistical method for combining the information of the individual indicators to produce lagging, coincidental, and leading composite indicators. There are also econometric-based models for formally combining information. These formal methods are briefly described next.

### *Statistical Modeling*

The purpose of statistical modeling is to identify a set of indicators that collectively represents all important aspects of an economy; to measure the indicators' statistical properties, such as their means, standard deviations, and correlation coefficients; and to combine this statistical information to produce composite indicators. The best-known such indicators are the Conference Board's composite lagging indicator (CLgI), composite coincidental indicator (CCI), and composite leading indicator (CLI) (Conference Board, 2001). The CLI is the most used as a predictor of the near future state of the U.S. economy.

CLI is a linear weighted-average value of 10 leading indicators of the U.S. economy whose assigned weights are standardized values of their relative variability, with standard deviation of an indicator used as a measure of variability. The higher the variability of an individual indicator (high standard deviation), the lower weight assigned to it in the composite. These weights are standardized to a total of 100% for all 10 indexes.

CLI values are produced monthly starting at a base value of 100 at a given month and year and adjusted for the changes in CLI every month after. Thus, an increase in CLI value over its previous month is considered a positive growth in the economy (i.e., increase in real GDP value), and a decrease in CLI value is considered a negative growth (i.e., a contraction). This base month and year is updated every few years. The latest update started in April 2008 by setting the CLI value for January 2004 equal to 100.

CLI's components and weights are regularly updated and revised to be current with the changing economy. Weight distribution is updated annually, reflecting the addition of new information used in statistical measures. However, more fundamental revisions happen less frequently and are made only when there is good evidence indicating a change in relative predictability of the individual leading indicators. This usually reflects a possible structural change in the economy.

The last such modification happened in 1996 when the Conference Board recognized that financial and money indicators were more accurate in predicting the near future state of the U.S. economy than the real economy indicators. A decision was made to replace three of the real economy indicators with two new money indicators. A revised distribution of the weights assigned a much higher total relative weight of about 65% to the money indicators and a much lower total weight of 35% to the real economy indicators. As a result of this revision, two money indicators of interest rate spread and money supply received over 60% of the total CLI weight.

This revision was troublesome because these two money indicators are also monetary policy instruments that, to a large extent, are controlled and influenced by the Federal Reserve Bank policy decisions. The implication is that monetary activities were to a large extent dominating CLI. In other words, CLI was less influenced by the activities of the real economy. This problem became obvious in the 2001 recession when CLI did not provide good forecasting information about the near future state of the economy. The Conference Board has, subsequently, made a revision to the definition of the interest rate spread in 2005, resulting in a reduced relative weight being assigned to it in CLI.

Criticisms of the statistical modeling of a composite indicator focus on its lack of underlying support in economic theory and on its not tracking a specific target similar to the one used by econometric modeling. However, from a practical viewpoint of being able to predict the near future state of the economy, especially in predicting and dating peaks and troughs in business cycles, the CLI has been very informative and accurate. In addition, CLI information is easy to understand and communicate to nontechnical consumers and beneficiaries (i.e., business decision makers) of such information.

### *Econometric Modeling*

Econometric modeling of composite indicators relies on cause-and-effect relationships supported by economic theory. Leading and coincidental indicators are modeled. Two approaches, factor-based modeling and Markov switching (MS) modeling, are used to model the composite coincidental indicator. For the development of the composite leading indicators, three approaches—vector autoregression (VAR)-based modeling, factor-based modeling, and MS-based modeling—are discussed next.

Econometric modeling of composite indicators is theoretical and complex, and requires a high degree of understanding of underlying technical issues. From the practical point of view of capturing and predicting the overall

behavior of the economy, the performance of these indicators is, at best, as good as that of indicators developed using statistical modeling.

The rationale behind factor-based modeling is that there are a number of common forces behind the movements of coincidental indicators. An econometric model that incorporates these common factors is developed to estimate the coincidental indicators (Sargent & Sims, 1977; Stock & Watson, 1989). The rationale behind MS modeling is the same with the exception of relaxing the assumption of a fixed dynamic behavior of the economy in a business cycle. In other words, as the economy goes through expansion and recession phases in a business cycle, the model parameters are changing or switching according to a probabilistic behavior of the Markov process (Hamilton, 1989). Justification behind MS modeling is that economic factors behave and interact differently in recession and expansion phases of a business cycle.

The rationales described here for factor-based and MS-based construction of composite coincidental indicators are also valid for the construction of the composite leading indicators (CLI). Because the main purpose of producing the CLI is predicting the near future state of the economy, the information generated from these models can be used to produce the probability of a turning point in an economy and to devise a decision rule for signaling a downturn point and a recession.

VAR is a linear model for establishing a relationship between leading indicators and coincidental indicators and their lagged values. Once a model is established, in a similar way to the other composite leading indicators' econometric models, its results can be used for business cycle predictions.

In addition to the approaches mentioned already, there are other mathematical approaches to developing leading indicators. One approach is incorporating a smooth transition to MS modeling. This approach assumes that the phases of the business cycle transition from one to another (i.e., expansion to recession) gradually rather than abruptly and suddenly. A second approach uses artificial neural networks to tap into a vast class of nonlinear models of the relationship between leading indicators and coincidental indicators. There is still another approach using binary variables in modeling expansion and recession phases of a business cycle and applying probit and logit methods for estimation and prediction. Massimiliano Marcellino (2006) is a good starting point for learning about modeling.

## Evaluation

A model is evaluated for its performance using in-sample data and out-of-sample data. In either case, this evaluation can be formal or descriptive. The purpose of the in-sample evaluation is to check for the fitness of the model to the sample data. Selection of a formal evaluation is a function of the forecasting methodology applied. For example, in cause-and-effect modeling of a target variable, such as the coincidental composite indicator as a linear function of leading indicators, standard statistical and econometric evaluative methods can be used. Formal methods for out-of-sample evaluation range from mean squared error (MSE) and mean absolute error (MAE), used if the target variable is a continuous variable such as a leading indicator, to quadratic probability score (QPS) and log probability score (LPS), used for predicting a discrete indicator such as an upturn or downturn in an economy (Diebold & Rudebusch, 1989; Winkler, 1969). A comprehensive treatment of forecast evaluation can be found in Kenneth D. West (2006).

A shortcoming of formal evaluation methods is that they are based on the information of the data series for the entire period of study, and every data point is given the same amount of weight in the analysis. In reality, data points around the peaks and troughs of a business cycle are more important for evaluating a turning point, and special attention should be given to them.

Descriptive evaluations of the models, which can be used for both in-sample and out-of-sample periods, are usually most effective in identifying the peaks and troughs of a business cycle. The goal is to measure how accurately and timely a model signals a turning point in an economy and to identify false alarms and missed signals. Descriptive evaluation methods are more informative in communicating the performance of a turning-point forecasting model.

A model that predicts a peak with a lead time is superior in its forecasting. Generally, the longer this lead time is, the better that model is. The number of consecutive downturn signals identified by a model is a measure of how persistent a model is in signaling a peak. The larger this value, the greater the likelihood that a peak will occur.

## Forecasting and Economic Policy Making

Economic climate is extremely uncertain and volatile during a slowdown, downturn, and peak period. The state of the economy, and the information about it, can change drastically in a very short period of time. A forecast is made based on the best information available at a given time, and an economic policy formulated using this forecast may become nonaccurate and ineffective to counterproductive, respectively, once the information has changed. This may create challenges for economists dealing with all stages of measuring economic indicators, of forecasting state of the economy, and of making policy recommendations. The contemporary state of the U.S. economy (2007–2009) has provided a clear example of this situation described as follows:

*2007 to 2008 Second Quarter:* Two major factors dominated the economic climate in this period.

1. Factor 1 is a substantial and continuing decline of the housing prices throughout the country after 5 years of major growth year after year. This has created two distinct problems for the economy with direct and, possibly, major effects on many economic activities. The first problem is a decline in the number of new housing construction permits and in the construction of the new houses. The

second problem, also known as the subprime mortgage rate crises, is the substantial readjustment increases of the variable mortgage rates as a result of the rise in the interest rates to the effect that a large number of these mortgages are not affordable to the homeowners. Nonaffordability of mortgage payments coupled with decline in housing prices have forced many banks to foreclose on people's houses. Because most of the mortgage contracts are securitized and sold in market to financial institutions throughout the world, it was unclear at this point how a large number of mortgage defaults could have an impact on these institutions in the coming months and years.

2. Factor 2 is the rapid increases in the commodity prices, from oil to soy beans, in the last few years, due to increase in demand from the emerging developing countries, which had put a substantial upward pressure on the prices of many goods and services from grocery products to plane tickets.

Confronting inflation in the economy is a primary policy objective of the Federal Reserve System for bringing stability in the economy. In formulating a monetary policy at that time, the Federal Reserve had to deal with the conflicting consequences of these two factors on the economy. On one side, it had to raise interest rates in order to confront inflation in the economy. On another side, it had to ease the supply of money in order to confront the economic slowdown consequences of a decline in housing construction and of instabilities in the financial institutions. The balancing policy actions of the Federal Reserve System in dealing with these problems were critical in preventing a peak in the U.S. economy in 2007 to 2008.

By summer 2008, all indications were that, of the two major factors impacting economic stability, the threat of the subprime mortgage rates effects was being diminished and inflation was going to be the main concern going forward. Therefore, what happened in the economy in late 2007 to early 2008 seemed most likely to be a slowdown, not a recession. The recommended policy action was an anti-inflationary policy of raising interest rates and reducing money supply.

*2008 to 2009:* Moving on to fall 2008, a sudden change occurred in the status of these two economic factors. It turned out that the impacts of the subprime mortgage rates on the financial institutions were so deep that they basically paralyzed the entire industry, requiring substantial governmental interventions, at the U.S. and global levels, to prevent it from collapsing. This brought slowdowns, downturns, and recessions to the countries around the world. In fact, in December 2008, NBER officially declared a recession for the United States starting December 2007. Given such a state of the economy globally, inflation did not turn out to be a concern, mainly due to a collapse in the commodity prices. The recommended policy was an expansionary policy of cutting interest rates and providing stimulus fiscal help.

## Economic Indicators and Forecasting in the Twenty-First Century

An economy, as a dynamic system, is continuously changing to incorporate technological innovations in the production processes and to satisfy the needs of an ever-evolving society. In the last 30 years, an outburst of technological innovations in the computer and in communication has moved the economy to be more global. This can clearly be seen in the formation of the global corporations in all industries, especially in financial, and in the outsourcing of production and services in the manufacturing and labor sectors. The global slowdown and recession of 2007 to 2009 has taught us that there is a strong interaction among the economic activities of all countries around the world. We are seeing that new economic powers are emerging in the developing countries, challenging the already well-established powers of the United States, the European Union, and Japan. These new powers are strong in both producing and consuming goods and services. At such a global setting, the economy of a country is very much dependent on global economic activities. This dependency needs to be incorporated in measuring economic parameters of a country, such as output, prices, employment, and so on. In addition, new parameters need to be identified and measured for the global economy. Further studies along these areas are expected in the years to come.

On the technical aspect of forecasting, one area that still needs substantial contribution is in applied forecast evaluation. Defining success in an organization and evaluating the performance of a forecast in that context is in critical need of development. For example, in predicting the overall state of the economy in the near future, is a forecast of a downturn that did not happen, due to the policy makers' effective and timely actions, a failure or a success? It can easily be argued that that forecast was a false alarm, thus, a failure. However, that predication has generated new information to alert the policy maker to act, thus it is a success. Developing objective evaluation methodologies of the performance of a forecast in the context of an organization's overall objective can be very helpful. For people interested in performing research on the technical areas of forecasting, a starting point is the excellent survey of Elliott et al. (2006).

## References and Further Readings

American Institute for Economic Research. (2008). *Understanding AIER's business-cycle research methodology.* Retrieved November 27, 2009, from http://www.aier.org/research/business-cycle

Burns, A. F., & Mitchell, W. C. (1946). *Measuring business cycles* (NBER Studies in Business Cycles No. 2). New York: National Bureau of Economic Research.

Conference Board. (2001). *Business cycle indicators handbook.* Available from http://www.conference-board.org/pdf_free/economics/bci/BCI-Handbook.pdf

Diebold, F. X., & Rudebusch, G. D. (1989). Scoring the leading indicators. *Journal of Business, 62*(3), 369–391.

Elliott, G., Granger, C., & Timmermann, A. (Eds.). (2006). *Handbook of economic forecasting* (Vol. 1). Amsterdam: Elsevier North-Holland.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica, 57,* 357–384.

Lahiri, K., & Moore, G. H. (1991). *Leading economic indicators: New approaches and forecasting records.* Cambridge, UK: Cambridge University Press.

Marcellino, M. (2006). Leading indicators. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 1, pp. 879–960). Amsterdam: Elsevier North-Holland.

Sargent, T. J., & Sims, C. A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. In C. Sims & C. W. Granger (Eds.), *New methods in business cycle research* (pp. 45–109). Minneapolis, MN: Federal Reserve Bank of Minneapolis.

Stock, J. H., & Watson, M. W. (1989). New indexes of coincident and leading economic indicators. In O. Blanchard & S. Fischer (Eds.), *NBER macroeconomics annual* (pp. 351–394). Cambridge: MIT Press.

West, K. D. (2006). Forecast evaluation. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 1, pp. 100–134). Amsterdam: Elsevier North-Holland.

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association, 64,* 1073–1078.

# 29

# Measuring and Evaluating Macroeconomic Performance

David Hudgins
*University of Oklahoma*

The performance of a nation's economy actualizes as both cause and effect. The phase of the economic cycle serves as a catalyst that permeates the decisions of individuals, business firms, and government policy makers on a continuous basis. It affects the prices and yields of financial assets, the spending decisions of consumers, corporate borrowing and investment, and choices available to general public. Yet the course of economic activity is also the result of the expectations and actions of its participants. Their collective behavior works in conjunction with the state of nature, which may include favorable conditions or unfavorable conditions, such as excessive droughts, flooding, or other disasters, to guide the economy along its trajectory.

Despite the importance of assessing economic performance, its measurement and evaluation are elusive. This is true for three main reasons. First, just as beauty is in the eye of the beholder, the state of the economy depends on the evaluator. There is no one universal criterion or natural law that determines the standard by which performance is judged. The choice of standard depends on a host of factors, some of which are the relative weighting of unemployment and inflation, the definitions of boom and recession, the rate of economic growth versus environmental degradation, and the national income distribution.

Second, macroeconomic performance deals with aggregated variables, such as overall economic output or national unemployment. These aggregated variables do not directly reveal all of the components that constitute the aggregates, such as the individual composition of the output, which means that many of the details are effectively hidden from policy makers. Third, the measurement of the variables themselves is imprecise. For example, there is no universal way to capture overall inflation. Even if accurate data on all prices could be accurately determined, the relative weighting scheme for determining which goods and services should contribute the most and least to the overall index is entirely up to the individual or organization constructing the index.

The primary reason for gathering and analyzing macroeconomic data is to aid in making decisions. Although the difficulties in assessing economic performance are formidable, there are some commonly accepted concepts that can be used. This chapter integrates these concepts to form a comprehensive approach.

## Theory and Concepts Relating to Economic Performance and the Business Cycle

*Macroeconomic performance* refers to the behavior of the entire economy. Economies follow a cyclical pattern, which is referred to as the economic cycle, or business cycle. The measurement over this cycle by macroeconomic data is the result of aggregation, which means combining many individual units into one overall economic unit. Although this has the downside of hiding the measurements of the individual components, it is necessary because the data must be combined to be useful. Because most markets tend to move in unison across the business cycle, much of the data

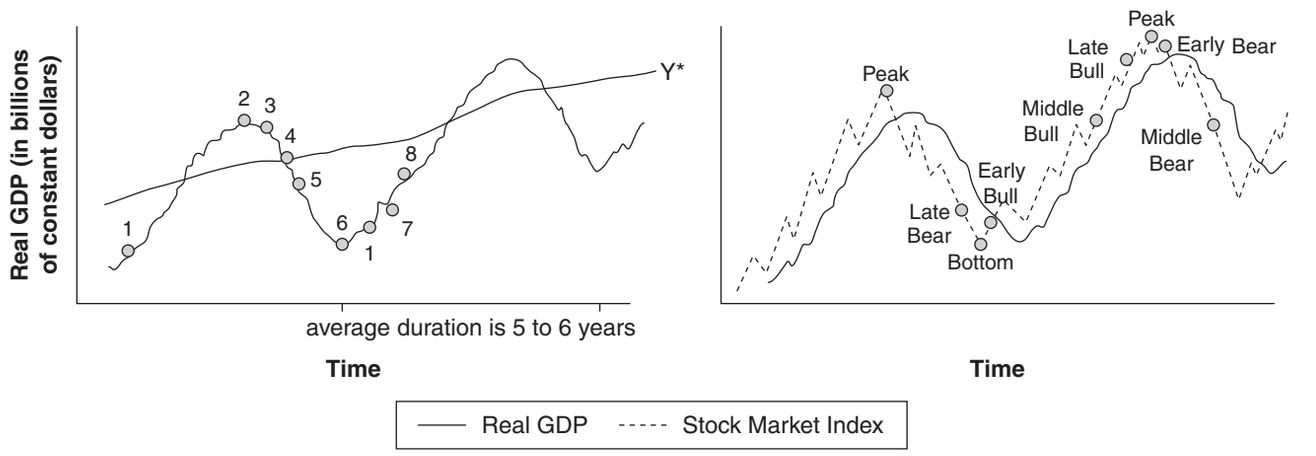


**Figure 29.1** The Business Cycle

NOTES: The first panel shows the phases of the business cycle along with Y* = potential GDP. The second panel shows the phases of the stock market cycle as a leading indicator of economic performance.

on individual markets can be relegated to the examination of economy-wide variables, which generally reflect an adequate signal of the phenomena whose measures are desired (Baumol & Blinder, 2008). The primary variables for measuring economic performance are the level and growth rate of national output (both overall and per capita), the inflation rate, and the unemployment rate.

The business cycle represents the tendency of economic activity to rise and fall over time in a cyclical pattern. It has four phases: (1) expansion (or recovery), (2) peak, (3) recession, and (4) trough (or bottom). When analyzing national output, the path of the business cycle can be traced out as in the first panel of Figure 29.1.

At Point 1, the expansion, or recovery, begins. During this phase, the growth rate of national output is high, while the national unemployment rate is falling. Once the level of economic activity reaches its peak at Point 2, it starts to decline, causing business inventories to rise around Point 3. At Point 4, a recession begins, eventually causing a high unemployment rate, a decline in interest rates, and a decline in the general level of prices, represented by Point 5. At Point 6, the economic trough, or bottom, is hit, and the economy starts on a new expansionary phase. At Point 7, consumption and investment rise, leading to a rise in interest rates and the general level of prices. Eventually, the economy reaches a new peak and will repeat the cycle again and again.

## Potential Gross Domestic Product

Notice that the cycle is not constant. Each successive peak tends to be higher than the last, so that the average level of economic output is constantly growing. This growth in the average level of economic output represents the upward long-run trend over time, and it is labeled as Y* in Figure 29.1. In the United States, this long-run trend average annual growth rate of real gross domestic product (GDP) is around 3%, with an average real per capita GDP growth rate of around 1.5% (*Economic Report of the President,* 2006; Peterson & Estenson, 1996). This is used as one benchmark for economic performance. An annual real GDP growth rate around 2% to 3% or higher reflects a strong economy, while a growth rate of less than 1% per year suggests poor economic performance.

This long-run trend value is of the upmost importance for several reasons. It represents what is equivalently referred to as potential GDP, full-employment GDP, or the natural level of output. The interpretation is that if the economy were not fluctuating due to cyclical factors, and thus, if it were always producing at its long-run sustainable rate of production, then it would be producing on this trend curve. Along this trend, the unemployment rate would consist of only what is referred to as frictional unemployment, meaning that all unemployment would result from only a small amount of normal job turnover. This trend unemployment rate is referred to as the natural rate of unemployment, or the nonaccelerating inflation rate of unemployment (NAIRU), and in the United States, it is generally assumed to lie between 4% and 5%. It is called the NAIRU because when the economy is operating on this long-run trend line, business firms are not producing at such a high level relative to current capacity that excess product demand and labor demand are causing inflationary pressures.

One final reason for the importance of this trend line of potential output is that it represents the material well-being of the economy's residents in the long run. Economist Paul Krugman (1990) stated that productivity is not everything, "but in the long run it is almost everything" (p. 9). For an economy's residents to sustain higher levels of per capita income and consumption, the economy must improve its capacity to produce output over time. Although this can be achieved in part by acquiring more of the factors of production such as capital, labor, and land, the primary means is by increasing labor productivity, which embodies both advances in technology and advancements in the skills and training of the workforce. This long-run trend curve Y* thus represents the growth rate of the economy's labor productivity, which in

turn determines the per capita consumption of the economy. Indeed, any macroeconomic policy that does not improve technology and labor productivity in the long run will not lead to any sustainable increase in the per capita income of its citizens.

One of the biggest obstacles in measuring and evaluating economic performance is that the graph in Figure 29.1 does not show all of the activity of the economy. The amalgam that is referred to as the economy does not evolve in cycles that are as clear-cut as those shown in Figure 29.1. Moreover, the economy never manifests the future values to a current observer, and the data that can be gathered, such as real GDP, generally has a lag involved in its collection and compilation. The trend curve for Y* is also only an estimation of the artificially generated concept called potential GDP and not an actually occurring phenomenon from which data can be collected.

Very few analysts ever agree on whether the economy is in an expansionary phase or a recessionary phase until the middle or even the end of the phase is reached. The term *recession* is properly defined as a period of time in which there is a marked decline in the general level of economic activity. It has commonly been defined as a decline in an economy's GDP during two consecutive quarters. The U.S. Department of Commerce's Bureau of Economic Analysis (BEA) has specifically stated that the widely used definition of a recession as a negative growth rate in GDP for two consecutive quarters is not correct. Instead, the BEA (2008) defers to the U.S. National Bureau of Economic Research (NBER) and states that

> recession is a monthly concept that takes account of a number of monthly indicators—such as employment, personal income, and industrial production—as well as quarterly GDP growth. Therefore, while negative GDP growth and recessions closely track each other, the consideration by the NBER of the monthly indicators, especially employment, means that the identification of a recession with two consecutive quarters of negative GDP growth does not always hold. (Business Cycle Dating Committee of the National Bureau of Economic Research, 2008)

The NBER has determined that during the period 1945 to 2001 in the United States, there were 10 business cycles with average duration from one trough to the next being 67 months.

## Cyclical Indicator Approach to Measurement of the Business Cycle

The business cycle is also alternatively assessed by the partially overlapping method referred to as the *cyclical indicator approach.* This approach analyzes what are called *economic indicators,* which are macroeconomic variables that are used to measure economic activity. These indicators serve as signals that represent the performance of various aspects of the economy. There are three types of economic indicators. Coincident indicators measure the current level of economic activity and turn at the same time as the business cycle turns. Leading indicators normally turn down before recessions start and turn up before expansions begin and can thus be used to forecast the direction in which the economic cycle is moving. Lagging indicators measure what the level of economic activity has been—they turn after the business cycle turns.

Rather than using just one indicator, the individual indicators are generally combined to form weighted composite indices. The Conference Board (2008) states, "These are constructed to summarize and reveal common turning point patterns in the data in a clearer and more convincing manner than any individual component, primarily because they smooth out some of the volatility of the individual components." The Conference Board, a nonprofit organization, now publishes the Index of Leading Economic Indicators (LEI) monthly for nine countries. Also called the Composite Index of Leading (Economic) Indicators, it was developed by the NBER and was formerly published by the BEA in the "chart book" section of the *Survey of Current Business.* It is supposed to forecast turns in the business cycle 3 to 6 months in advance but is often unreliable. Economist Paul Samuelson (1966) famously stated that "Wall Street indexes predicted 9 of the last 5 recessions" (p. 92), and this statement is widely applied as a criticism of the LEI as a forecasting or measuring tool. The current components of the LEI, in decreasing order of their relative weights, are as follows:

1. Interest rate spread: 10-year Treasury bond (T-bond) rate minus federal (fed) funds rate. This is a measure of the slope of the yield curve. A T-bond is a long-term coupon-bearing debt issued by the U.S. government with a maturity that is generally more than 7 years. The interest is paid semiannually and could be exempted from the local and state tax. The fed funds rate is set by the U.S. Federal Reserve System (the Fed) and is the overnight interest rate at which banks lend the federal funds to other depository institutions.
2. Real M2 money supply. M2, a proxy of money supply, is the summary of physical currency, checking and saving accounts, small denomination consumer certificates of deposit, and noninstitutional money market mutual funds.
3. Average weekly hours of production workers in manufacturing. These data are issued monthly by the U.S. Department of Labor's Bureau of Labor Statistics (BLS) (http://www.bls.gov). Both seasonally adjusted and nonseasonally adjusted data are released.
4. Manufacturer's new orders for consumer goods and materials, adjusted for inflation. These data are issued monthly by the U.S. Department of Commerce's Census Bureau. Both seasonally adjusted and nonseasonally adjusted data are released. New orders are the totals for that period, and these data are adjusted based on the calendar month variations and

trading days. Data for the semiconductor industry are not included.
5. Stock prices, measured by the Standard & Poor's 500 Index (S&P 500). The S&P 500 is the most commonly used benchmark for the U.S. equity market and includes the top 500 large-cap (large capitalization) companies, which are from different industries, such as energy, materials, industrials, consumer discretionary, consumer staples, health care, financials, information technology, telecommunications services, and utilities.
6. Fraction of vendors facing slower deliveries from suppliers, a component of the National Association of Purchasing Managers index.
7. Average weekly initial claims for state unemployment insurance (inverted). This index is issued weekly by the U.S. Department of Labor. Both seasonally adjusted and nonseasonally adjusted data are released.
8. New building permits issued. This index is issued monthly by the U.S. Department of Commerce's Census Bureau. Both seasonally adjusted and nonseasonally adjusted data are released.
9. Index of Consumer Expectations. These data are issued monthly by the University of Michigan and reflect American consumers' subjective expectation for the current and future (6 months) economic situation. This index also plays an important role in the decisions of business firms and in the public polices of government.
10. Manufacturers' new orders for nondefense capital goods, excluding aircraft. These data are issued monthly by the U.S. Department of Commerce's Census Bureau. Both seasonally adjusted and nonseasonally adjusted data are released.

The series in the LEI that are based on the Conference Board estimates are manufacturers' new orders for consumer goods and materials, manufacturers' new orders for nondefense capital goods, and the personal consumption expenditure used to deflate the money supply.

The Conference Board also publishes an Index of Coincident Economic Indicators and an Index of Lagging Economic Indicators. The four components of the Index of Coincident Economic Indicators are as follows:

1. Personal income less transfer payments
2. Manufacturing and trade sales
3. Industrial production
4. Employees on nonagricultural payrolls

The components of the Index of Lagging Economic Indicators include the following:

1. The average duration of unemployment (inverted)
2. Change in the consumer price index (CPI) for services
3. Ratio of consumer installment credit to personal income
4. Change in labor cost per unit of output
5. Commercial and industrial loans outstanding
6. Average prime rate charged by banks
7. Ratio of manufacturing and trade inventories to sales

Two other indices are also used. The ratio of the index of coincident to lagging indicators is used by some forecasters as a leading indicator. Diffusion indices, generally, analyze the degree to which the individual components of an index move with the direction of the overall index. Each diffusion index published by the Conference Board measures the proportion of the components that contribute positively to the overall index.

## Measuring the Economy Through Stock Market Cycles

The path of the stock market cycle generally precedes the business cycle, and thus, the performance of the stock market is a leading indicator. As previously mentioned, stock prices as measured by the S&P 500 are the fifth indicator in the LEI index. Because the prices of equities are determined by both the perceived underlying company fundamentals and the integration of all investor expectations, the stock market performance is symptomatic of the overall direction of the economy. A *bull market* is a sustained period of time when the broad indices of stock prices are rising, or advancing. A *bear market* is a sustained period of time when the broad indices of stock prices are falling, or declining.

As shown in the second graph in Figure 29.1, the early bull market precedes the general economic cycle trough, and the early bear market generally precedes the peak of the economic cycle. This fact provides a chief reason why the stock market is watched so closely by analysts: It signals the direction that the overall economy is likely to traverse in the near future. It is also the reason why it is so important to determine when the stock market hits a bottom after a bear market. Once the bottom is reached, the overall economy begins to pull out of a recession and begins a recovery phase.

The behavior of the stock market also partially drives overall economic performance. Standard economic analysis assumes that private aggregate consumption spending, which is the largest component of a nation's aggregate demand (about 70% in the United States), depends on both per capita income and the value of real wealth in consumers' portfolios. If stocks are in a bear market, then individuals will find that the balance in their retirement accounts and other investment portfolio accounts has dropped. As a result, consumption spending, consumer and business confidence, and business investment are all likely to fall, thus leading to an economic downturn.

Stocks tend to fall into three broad categories. They either move with the business cycles (cyclical stocks), resist these cycles (defensive stocks), or assume a long-term trend of growth despite market ups and downs (growth stocks). *Cyclical stocks* are stocks whose prices vary with anticipated company earnings, which in turn move together with the swings in the business cycle. In general, they include steel, cement, paper, aluminum, machinery, machine tools, airlines, railroads and railroad equipment, automobiles, and other industrial groups. *Defensive stocks* are stocks that present relatively low risk

in recessions because they represent corporations whose products or services are essential to consumer needs and are sufficiently low priced to be demanded at all stages of the business cycle. They include foods, tobacco, utilities, and household products. *Growth stocks* are stocks of corporations whose sales and earnings have increased faster than the general economy over a secular time span (more than one complete business cycle). They may come from all sectors, especially technology.

When a recession appears likely, many investors will shift their portfolio holdings toward defensive stocks. If the recession appears to be especially severe, investors will shift mostly out of stocks and into cash, money market assets, and bonds. When interest rates (and anticipated stock prices) have approached a bottom, and when the outlook is for economic recovery, then investors generally shift toward cyclical and growth stocks. Hence, the overall composition of financial portfolios also reflects economic conditions as well as consumers' expectations about the path that the economy will take in the near future.

One of the key determinants of the price movement of a stock is its quarterly earnings per share (EPS). If the EPS shows strong growth as compared to the same quarter in the previous year for an increasing number of stocks in a recurring pattern over several quarters, then corporations tend to show strong profits and growth. If the EPS is increasing for a large percentage of corporations, then this signals that the economy is strong, thus creating a favorable overall climate for increases in stock prices and an economic boom. A lackluster growth or decline in key corporate sectors or in a wide-ranging group of firms indicates weak economic fundamentals and an impending cyclical downturn. Thus, corporate earnings reports are watched closely by market analysts.

## Measuring the Pattern and Robustness of the Economic Cycle

In the early 1930s, the world was undeniably in the grip of the Great Depression. In 1931, U.S. President Herbert Hoover assigned economist Edward R. Dewey, chief economic analyst for the U.S. Department of Commerce, the task of uncovering the causes of the economic crash and the misery that lay behind this period of suffering. Dewey formed and became the president of the Foundation for the Study of Cycles in 1942, and this now defunct organization was superseded by the Cycles Research Institute (CRI; http://www.cyclesresearchinstitute.org) in 2004.

In over 30 years of research on cycles, Dewey and his associates studied empirical data from a huge variety of natural and human cycles, including insects, fish, mammals, man, and other natural phenomena, which were all found to be connected to sunspot cycles. Some examples included caterpillars in New Jersey, lynx and coyote abundance in Canada, salmon abundance in Canada and England, heart disease in New England, ozone content in London and Paris, pig iron prices, steel production, cigarette production, Goodyear tire and rubber sales, railroad stock prices, rainfall in London, worldwide precipitation, and the average yield of chief crops in Illinois. Dewey found cycles with periods ranging from months to hundreds of years, and several thousand cycles were recorded. One of the main cycles was one with a period of 9 2/3 years (Dewey, 1967).

Based on Dewey's analysis, one other economic assessment tool has been acquired—that of indicators from other disciplines. Because there is both theoretical and empirical evidence that economic cycles are correlated with other human and physical phenomena, the performance of the economy can also be measured by indicators from these other fields. This is obvious in a direct sense, because agricultural production depends on natural weather cycles and economic productivity depends on the stability of the social and political systems; for example, countries being devastated by war will not have strong economic performance.

Dewey concluded that four of the empirical laws of cycles are as follows:

1. Common cycle periods appear in many seemingly unrelated disciplines.
2. Synchrony is the observation that cycles of the same period often have the same phase.
3. Cycles' harmonic ratios is the observation that the common cycle periods are related by the ratios 2, 3 and their products.
4. Cycles outside the earth are often related to cycles on earth. (CRI, n.d.)

To optimally assess the performance of the economy, the driving influences and the responses must be adequately understood. Dewey concentrated on *forced oscillations,* which are externally caused rhythms that are patterns where the rhythm in the economic and other cycles is the direct result of a rhythm in some other phenomenon. Dewey (1967) concluded,

> Insofar as cycles are meaningful, all science that has been developed in the absence of cycle knowledge is inadequate and partial. Thus, if cyclic forces are real, any theory of economics, or sociology, or history, or medicine, or climatology that ignores non-chance rhythms is manifestly incomplete . . . as medicine was before the discovery of germs.

Ongoing comprehensive studies across diverse phenomena provide an increasingly clearer picture of the state of the economy, especially with the current rapid advances in data analysis, computing, and communications technology.

### Schumpeter's Waveform Analysis of the Economic Cycles

Further insight into the understanding of the measurement of the economy and its cycles was added by the well-known and respected economist Joseph Schumpeter. Similar to Dewey, Schumpeter (1938) used a variety of

empirical evidence to develop a cyclical waveform model of economic behavior. This model consisted of a combination of three cycles. The first was Nikolai Kondratieff's long-wave cycle of 50 to 60 years. The second was French medical doctor Clement Juglar's 9 to 10 intermediate-term cycle, which is in accordance with Dewey's 9 2/3-year cycle. The third was economist Joseph Kitchin's short-term minor cycle of 40 months.

Unlike Dewey, Schumpeter (1938) viewed his analysis as being very limited as a useful policy tool. He stated,

> No claims are made for our three cycle scheme except that it is a useful descriptive or illustrative device. Using it, however, we in fact got *ex visu* of 1929, a "forecast" of a serious depression embodied in the formula: coincidence of the depression phase of all three cycles. (p. 174)

This illustrates the principle of resonance in the business cycle models: When the long-, intermediate-, and short-run wave cycles all coincide in a downturn, the economy is likely to suffer an extreme depression. On the other hand, coincidence of all of the upward short-, intermediate-, and long-run cycles creates explosive growth while the economy transitions into a new, higher phase. For several decades, many economists had rejected the ideas of Dewey, Schumpeter, and others who studied these cyclic patterns. Current world developments and newfound computational tools have eclipsed this narrow-minded stance, having shown that this broad, multidisciplinary approach serves as a useful tool in both the measurement and forecasting of economic performance at the macro and micro levels.

### The New Keynesian and Real Business Cycle Views on Evaluating Performance

There are two directly contrasting views on how the economy's performance should be treated as it reacts to stimuli, or "shocks," over the course of the business cycle (Mankiw, 1989; Peterson & Estenson, 1996; Plosser, 1989). The real business cycle (RBC) approach extends the neoclassical framework and suggests that when supply shocks such as technology changes occur, or when demand shocks such as a change in consumer expectations occur, economic participants always act optimally so as to maximize their welfare. As a result, active fiscal and monetary policy have no role because all markets always clear and are in continuous equilibrium. The new Keynesian view is built on the foundation that the reason for recessionary periods is due to market failure, and thus, the economy frequently performs at suboptimal levels away from its efficient full-employment equilibrium.

When looking at the Great Depression and other major recessionary periods worldwide, especially when considering the corruption-induced contributions in the United States such as the saving and loan crisis of the 1980s, the dot-com scandals, and the subprime real estate and financial crises of the new millennium, it is not plausible for even the most ardent proponent of RBC to argue that the economy has continually followed an optimal equilibrium adjustment path. Nonetheless, just stating that the economy is following a nonequilibrium path and attempting to measure the hypothetical "income gap" is not a full solution to the measurement problem. The measurements must be augmented with the insights of the more robust analyses of the type used by Dewey and Schumpeter in order to study the nature of the shocks themselves, which are never "external" but rather part of the interconnected plethora of processes that jointly determine the level of economic performance that is to be evaluated as it unfolds over time.

## The Primary Variables and Data Used to Assess Macroeconomic Performance

National output, inflation, and unemployment are the undisputed "three kings" of the data used to gauge economic conditions. A nation's output is virtually equivalently measured by three primary measures of economic production: gross national product (GNP), which is based on ownership of the factors of production; GDP, which is based on geographical borders; and gross national income (GNI), which is based on the income earned by factors of production. Quarterly and annual data for all of these in the United States are published by the BEA in the National Income and Product Accounts, which the BEA considers to be the cornerstone of all its statistics. Some of these data are also reprinted annually in the tables section of the *Economic Report of the President.*

GDP is the sum of the money values of all final goods and services produced by the economy during a specified period of time. It includes all production inside the country's geographical borders. Nominal GDP (NGDP) is calculated by valuing the outputs at current prices, whereas real GDP (RGDP) is calculated by valuing all outputs at the prices that prevailed in some base year, which is currently the year 2000. Therefore, RGDP directly measures the quantities produced, and is a far better measure of changes in national production than is nominal GDP.

However, RGDP is a secondarily constructed measure, which can be seen by viewing the manner in which it is obtained. In the United States, for the year 2001, nominal GDP was $10,128 billion. The GDP implicit price deflator (IPD), which is the quarterly published index of overall prices, had a value of 100 in the year 2000 because it is the currently used base year. The IPD price index value for the year 2001 was 102.399, meaning that the overall level of prices was 2.399% higher in the year 2001 than in 2000. This percentage change,

which measures the overall inflation rate, is found by the following formula:

$$\begin{aligned}\text{percentage change} &= \frac{\text{new} - \text{old}}{\text{old}} \\ &= \frac{102.399 - 100}{100} \\ &= 2.399\% \\ &= \text{inflation rate for 2001.}\end{aligned}$$

RGDP is found by dividing nominal GDP by the IPD index for the corresponding year. Before this calculation is made, however, the IPD index must be taken out of index number form by dividing it by 100. In the previous example, the transformed IPD value equals 1 for the base year 2000 and 1.02399 for the year 2001. In the year 2000, nominal GDP was \$9,817 billion, so

$$\begin{aligned}\text{real GDP}_{2000} &= \frac{\text{nominal GDP}}{\text{transformed IPD}} \\ &= \frac{\$9,817 \text{ billion}}{1} \\ &= \$9,817 \text{ billion.}\end{aligned}$$

Note that values of nominal and real GDP are always equal for the base year. In 2001,

$$\begin{aligned}\text{real GDP}_{2001} &= \frac{\$10,128 \text{ billion}}{1.02399} \\ &= \$9,890.72 \text{ billion of constant} \\ &\quad (\text{year } 2000).\end{aligned}$$

The growth rate of the economy is measured by the percentage change in the RGDP. The growth rate for the year 2001 is as follows:

$$\begin{aligned}&\frac{\$9,890.72 \text{ billion} - \$9,817 \text{ billion}}{\$9,817 \text{ billion}} \\ &= 0.00751 = 0.751\%.\end{aligned}$$

This is the same growth rate that can be compared to the 2% to 3% annual growth of the trend curve given by Y* in Figure 29.1. Because 0.751% is less than even 1% annual growth, this year can be evaluated as a slow growth year.

Both internally and in cross-country comparisons, the most widely used measure of economic performance regarding the level and change in a country's economic well-being is its per capita RGDP, or RGDP per person. Per capita RGDP represents the purchasing power of the average citizen's annual income earnings. It is found by using the following formula, where the year 2001 is again used as a continuing example:

$$\begin{aligned}\text{per capita RGDP}_{2000} &= \frac{\text{RGDP}}{\text{population}} \\ &= \frac{\$9,817 \text{ billion}}{282{,}403{,}000 \text{ persons}} \\ &= \$34{,}762.38/\text{person.}\end{aligned}$$

$$\begin{aligned}\text{per capita RGDP}_{2001} &= \frac{\text{RGDP}}{\text{population}} \\ &= \frac{\$9,890.72 \text{ billion}}{285{,}335{,}000 \text{ persons}} \\ &= \$34{,}663.54/\text{person.}\end{aligned}$$

The growth rate for per capita GDP in the year 2001 was

$$\begin{aligned}&\frac{\$34,663.54 - \$34,762.38}{\$34,762.38} \\ &= 0.28\%.\end{aligned}$$

This small negative growth rate means that the average citizen encountered a fall in the real income level and would have only been able to purchase a slightly smaller quantity of goods and services than in the previous year. In years where the economy is growing according to its long-run full employment equilibrium trend growth rate, this per capita increase would be around 2% to 3%.

Real per capita GDP (or GNP) is not a full measure of a nation's economic well-being and does not provide precisely accurate cross-country comparisons because of the following reasons (Baumol & Blinder, 2008; Nafziger, 2006). First, for the most part, only goods and services that pass through organized markets are counted in the GDP calculations. International GDP comparisons are vastly misleading when countries differ greatly in the fraction of economic activity that each conducts in organized markets. The underground economy, which consists of many cash and barter transactions and illegal activities, is not counted, neither are the services of housewives or househusbands.

Another shortcoming is that GDP places no value on leisure, so GDP would understate economic well-being in this respect. Negative and destructive types of goods and those that are produced as a result of a negative stimulus also get counted in GDP, so GDP could overstate economic well-being in this respect. For example, although production might increase to clean up after an earthquake or to fight a war, the citizens would be better off without these events. Ecological costs are not netted out of GDP, either, even though environmentally adjusted measures have been proposed.

GDP is understated for developing countries, where household size is substantially larger than that in developed countries, resulting in household scale economies. Two people can live more cheaply together than separately. India's household size is 5.2 compared to the United States' 2.6. Moreover, a larger percentage of the average Indian household consists of children, who consume less food and resources than adults. Equivalent adult, equivalent household (EAEH) adjusts income based on household size and children. India's per capita income is about 10% higher when it is EAEH adjusted, and Africa's EAEH adjustment for per capita income is more than 10% because

its population growth rate and average household size are larger than India's (Nafziger, 2006).

GDP includes only what are called final goods, which are those that are purchased by their final users. It does not include intermediate goods purchased for resale or for use in producing another good; otherwise, there would be multiple counting of these intermediate goods. It has been argued that many of the final goods produced in developed countries, such as Western executives' business suits, smog eradication expenditures, and part of defense spending, should in fact be classified as intermediate goods (Nafziger, 2006). As a result, GDP may be overstated for developed countries because a number of items included in their national output are intermediate goods, reflecting the costs of producing or guarding income.

For international comparisons, the exchange rate used to convert GDP in local currency units into U.S. dollars, if market clearing, is based on the relative prices of internationally traded goods (and not on purchasing power). However, GDP is understated for developing countries because many of their cheap, labor-intensive, unstandardized goods and services have no impact on the exchange rate, because they are not traded. Many of the necessities of life, such as food—for example, cheap rice is the primary food staple in India—are very low priced in dollar terms. GDP is overstated for countries, usually developing countries, where the price of foreign exchange is less than a market-clearing price. On balance, however, the other adjustments outweigh this effect so that income differences between rich and poor countries tend to be overstated.

GDP growth is heavily weighted by the income shares of the rich. A given growth rate for the rich has much more impact on total growth than the same growth rate for the poor. In India, which has moderate income inequality, the upper 50% of income recipients receive about 70% ($350 billion) of the GDP, and the lower 50% receive about 25% ($150 billion) of total GDP ($500 billion). A growth of 10% ($35 billion) in income for the top half results in 7% total growth, but a 10% ($15 billion) income growth for the bottom half is only 3% aggregate growth. However, the 10% growth for the lower half does far more to reduce poverty than the same growth for the upper half.

## The Twin Evils of Inflation and Unemployment

Inflation and unemployment are known as the twin evils of macroeconomics. One of the key macroeconomic goals for every country is to have an inflation rate that is very stable and close to zero. *Inflation* is defined as a sustained increase in the general price level. As previously discussed, the inflation rate is measured by the percentage growth in the price index. There are three primary price indices used to measure inflation. The GDP IPD discussed previously is the most comprehensive and is published quarterly by the BEA. It is a type of Paasche aggregate price index, which means that it uses current year quantities as weights for its calculation. Thus, the IPD index shows how much the quantity of goods constituting GDP produced in a given year would have cost if those same goods would have been produced in the base year. The weights for the index thus change in every period.

The other two indices are both published by the U.S. Department of Labor's BLS. The producer price index (PPI) measures the selling prices received by domestic producers for their output. The PPI targets the entire marketed output for U.S. producers and thus excludes imports. The prices included in the PPI are from the first commercial transaction for many products and some services.

The most commonly quoted inflation measure is the consumer price index (CPI), which measures the prices paid by urban consumers for a representative basket of goods. Whereas the primary use of the PPI is to deflate revenue streams in order to measure real growth in output, the primary use of the CPI is to adjust income and expenditure streams for changes in the cost of living. The CPI index is a type of moving Laspeyres aggregate price index, which uses the base year quantities as weights. This index shows how much a basket of goods produced in the base year (say, the year 2000) would have cost in any given year of interest. The BLS publishes three major versions of the CPI. The CPI for all urban consumers (CPI-U) is a broader index than the original CPI index, and it is the most used version quoted in the media. Individual income tax parameters and treasury inflation-protected securities returns are based on all the items CPI-U. In 1978, the CPI-U was created and the original CPI was renamed as the CPI for urban wage earners and clerical workers (CPI-W). Social Security and federal retirement benefits are updated each year for inflation by the CPI-W.

The validity and optimality methods of assessing inflation through the methods employed by the BLS in computing the CPI are controversial (Greenlees & McClelland, 2008; Williams, 2006). Former chairman of the Fed Alan Greenspan and Michael Boskin put forth a substitution effect argument that stated that the CPI would overstate inflation if consumers substituted hamburger for steak when the price of steak became more expensive. The CPI does not use a variable basket of goods that allows this type of substitution; however, the original arithmetic weighting scheme was replaced by a geometric weighting scheme that now automatically gives a lower weight to CPI components that are rising in price and a greater weight to items whose prices are falling. The net effect has been a reduction in the annually reported CPI by 2.7% compared to the original arithmetic weighting scheme, and a reduction in the Social Security annual cost of living adjustment by more than a third when compounding the results from the 1990s to the present (Williams, 2006). The BLS defends the geometric weighting scheme in its formula, noting that it is widely used by worldwide statistical agencies and that it is recommended by the International Monetary Fund and the Statistical Office of the European Communities.

Another controversial aspect of the CPI formula is hedonic quality adjustment, which is one of the methods used to determine what portion of a price difference is viewed by consumers as reflecting quality differences. The BLS (Greenlees & McClelland, 2008) defines this as "a statistical procedure in which the market valuation of a feature is estimated by comparing the prices of items with and without that feature." For example, when the price of a television rises, the hedonic adjustment would assume that part of the price increase was due to an increase in the quality and part of the increase was due to inflation, so that the CPI would view the new, higher price as reflecting the higher quality or added features. This adjustment is necessary to reflect both improvements and drops in the quality of goods; however, the amount of adjustment and the range of goods for which it is applied remain subjective. If the quality adjustments are weighted too heavily, then the CPI will understate the true underlying inflation rate experienced by consumers. The chained CPI-U (C-CPI-U), introduced during the second Bush administration as an alternate CPI measure, attempts to directly measure the substitution effect, rather than an approximation based on a geometrically weighting scheme.

A final controversy lies in the short-term and seasonal adjustment procedures. When the index is seasonally adjusted for monthly repeating seasonal price swings, intervention analysis dampens the volatility of extreme fluctuations in fuel prices, especially price increases, which results in the CPI reflecting a lower inflation rate than that which is actually experienced. This problem is related to the controversial concept of the core rate of inflation (net of food and energy), which also seeks to remove the short-term (1- or 2-month) volatility resulting from extremely large fluctuations in food and energy prices. This is always be a misleading inflation gauge because food and energy account for about 23% of consumer spending as weighted in the CPI (Williams, 2006).

The third part of the trilogy of measures of performance is the unemployment rate. High unemployment has been linked to higher incidence of certain types of crimes, psychological disorders, divorces, and suicides (Baumol & Blinder, 2008; Mankiw, 2006). Unemployment may also cause a loss of job skills (or a decrease in their accumulation).

The BLS measures unemployment by surveying about 60,000 households every month. Workers are classified as being employed, unemployed, or not in the labor force. The labor force refers to the number of people holding or seeking jobs, so that the labor force = employed + unemployed. The unemployment rate is defined as the percentage of the labor force that is unemployed. For the United States in the year 2001,

$$\text{labor force} = 135.1 \text{ million (employed)} + 6.7 \text{ million (unemployed)} = 141.8 \text{ million}$$

$$\text{u} = \text{unemployment rate} = \frac{\text{unemployed}}{\text{labor force}} = \frac{6.7 \text{ million}}{141.8 \text{ million}} = 0.047 = 4.7\%.$$

Although both civilian and overall unemployment rates are computed, it is the civilian unemployment rate that is usually quoted. Unemployment rates are also computed for more narrowly defined groups—blacks, whites, men, women, and so on.

The stated unemployment rate figures actually understate the true underlying unemployment problem by ignoring discouraged workers and underemployed workers. When individuals desire to work full-time but are currently employed only part-time, they are counted as employed. The unemployment rate ignores the underemployment of these workers. The unemployment rate also does not count discouraged workers, who are unemployed persons that give up looking for work and are therefore no longer counted as being in the labor force.

Conversely, the unemployment rate overstates the true unemployment burden as the result of what is referred to as the added worker effect, which occurs when nonworking spouses or teenagers enter the labor force as the result of a primary household income earner becoming unemployed. The added worker effect smooths family income over the business cycle. In the United States, most spells of unemployment are short, and most of the unemployment observed at any given time is long-term. Thus, most of the economy's unemployment problem is attributable to the relatively few workers who are jobless for long periods of time (Mankiw, 2006).

There are four types of unemployment. *Frictional unemployment* is due to normal turnover in the labor market. It includes people who are temporarily between jobs because they are moving or changing jobs (or occupations). *Structural unemployment* occurs when workers lose their jobs because they have been displaced by automation, because their skills are no longer in demand, because of sectoral shift, or other similar reasons. *Seasonal unemployment* results when employment is of a seasonal nature, such as with lifeguarding, ski instructors, and construction work. These three are sometimes classified together as frictionally unemployed.

The last type is the one negative type, which results from poor economic performance. *Cyclical unemployment* results due to a decline in the economy's total production. It rises during recessions and falls during booms. The Employment Act of 1946 committed the U.S. government to help maintain low unemployment.

*Full employment* is a situation where no cyclical unemployment and minimal structural unemployment exists. The natural rate of unemployment is the unemployment rate that corresponds to full employment, where there is no cyclical unemployment. It is the normal rate of unemployment around which the actual unemployment rate

fluctuates over time. As explained previously, it is also called the NAIRU, and in the United States it fluctuates between 4% and 5%. So, even at full employment, some unemployment exists.

## Conclusion

This analysis has presented the data sources and methods for measuring the macroeconomic variables that determine economic performance. Over the course of the business cycle, the level of economic activity generates a slew of indicators as it waxes and wanes across the time series that are to be evaluated.

Many macroeconomic models that measure economic performance define a performance or criterion index that allows the policy maker to choose a compromising trade-off between high inflation and high unemployment (which corresponds to low output growth, especially in the short run). This inflation–unemployment trade-off is referred to as the Phillips curve relationship. Evaluation of economic performance always depends on the relative weights assigned to price stability, unemployment, and output growth. There is no global positioning system to define where the economy is and which path will most directly lead it to a desired future destination. As it proceeds, the economy's path will be evaluated differently by each individual person based on the perceived influence of its entirety on that person's experience.

## References and Further Readings


Baumol, W., & Blinder, A. (2008). *Economics: Principles and policy* (11th ed.). Mason, OH: South-Western.

Bureau of Economic Analysis. (2008). *Glossary.* Retrieved January 11, 2009, from http://www.bea.gov/glossary/glossary.cfm?letter=R

Business Cycle Dating Committee of the National Bureau of Economic Research. (2008, January). *The NBER's recession dating procedure.* Retrieved January 11, 2009, from http://www.nber.org/cycles/jan08bcdc_memo.html

Conference Board. (2008, August). *U.S. leading economic indicators and related composite indexes for July 2008.* Retrieved January 11, 2009, from http://www.conference-board.org/pdf_free/economics/bci/LEI0808.pdf

Conference Board. (n.d.-a). *Global business cycle indicators.* Retrieved January 11, 2009, from http://www.conference-board.org/economics/bci

Conference Board. (n.d.-b). *How to compute diffusion indices.* Retrieved January 11, 2009, from http://www.conference board.org/economics/bci/di_computation.cfm

Cycles Research Institute. (n.d.). *Edward R. Dewey.* Retrieved January 11, 2009, from http://www.cyclesresearchinstitute.org/dewey.html

Dewey, E. R. (1967, July). The case for cycles. *Cycles.* Retrieved January 11, 2009, from http://www.cyclesresearchinstitute.org/dewey/case_for_cycles.pdf

Dewey, E. R., & Dakin, E. F. (1964). *Cycles: The science of prediction.* Albuquerque, NM: Foundation for the Study of Cycles. (Original work pulished 1947)

*Economic report of the president. Transmitted to the Congress February 2006, together with the annual report of the Council of Economic Advisers.* (2006). Washington, DC: Government Printing Office.

Greenlees, J., & McClelland, R. (2008, August). *Common misconceptions about the consumer price index: Questions and answers.* Retrieved January 11, 2009, from http://www.bls.gov/cpi/cpiqa.htm

Krugman, P. R. (1990). *The age of diminished expectations.* Cambridge: MIT Press.

Mankiw, N. G. (1989). Real business cycles: A new Keynesian perspective. *Journal of Economic Perspectives, 3*(3), 17–90.

Mankiw, N. G. (2006). *Macroeconomics* (6th ed.). New York: Worth.

Nafziger, E. W. (2006). *Economic development* (4th ed.). Cambridge, UK: Cambridge University Press.

National Bureau of Economic Research. (n.d.). *Business cycle expansions and contractions.* Retrieved January 11, 2009, from http://wwwdev.nber.org/cycles/cyclesmain.html

Peterson, W., & Estenson, P. (1996). *Income, employment, economic growth* (8th ed.). New York: W. W. Norton.

Plosser, C. I. (1989). Understanding real business cycles. *Journal of Economic Perspectives, 3*(3), 51–77.

Samuelson, P. (1966, September 19). Science and stocks. *Newsweek,* p. 92.

Schumpeter, J. (1938). *Business cycles* (Vols. 1–2). New York: McGraw-Hill.

Williams, W. J. (2006). *The consumer price index: Government economic reports: Things you've suspected but were afraid to ask!* Retrieved January 11, 2009, from http://www.shadowstats.com/article/56

Zarnowitz, V. (2008, July). *Present conditions of the U.S. economy in historical perspective.* Retrieved January 11, 2009, from http://www.conference-board.org/pdf_free/economics/PresentUSEconomy.pdf

# 30

# Macroeconomic Models

Christopher J. Niggle
*University of Redlands*

The term *macroeconomics* refers to study of the behavior of an economy as a whole or as a system; the phenomena explained are (1) the short-run level of economic activity—the levels of national output, income, and employment; (2) the causes of short-run fluctuation in economic activity (business cycles); and (3) the long-run growth rate of an economy. This chapter focuses on the first two aspects of macroeconomics. The models are presented in an approximate chronological order; the chapter's organizing theme is that modern macroeconomic models can be seen as based on one of two competing "visions" of the economy: (1) The economy is seen as stable, with strong market forces pushing it toward an equilibrium level consistent with full employment of labor and capital (as in the classical and new classical models), or (2) it is seen as an unstable system that grows through time in a boom–bust pattern, with its normal state being less than full employment and so less-than-potential output being produced (as in Keynes's and the Keynesians' models).

## The Beginning: Keynes's Critique of Classical Economics

Macroeconomics as a distinct field within economics emerged in the late 1930s as a response to John Maynard Keynes's *General Theory of Employment, Interest and Money* (1936/1973, referred to subsequently as GT). Keynes contrasted his views on the causes of depressions and persistent involuntary unemployment with those of his predecessors, whom he termed the *classical economists.* In Keynes's view, these economists assumed that the normal condition for a market economy is one of capital-accumulation-fueled growth with full employment of labor and capital, and that periods of high unemployment were rare and temporary deviations from the norm. Keynes wrote that this assumption is empirically incorrect because economies frequently experienced prolonged periods of high unemployment and below-potential output (recessions or depressions). He presents his model by developing a critique of three dimensions of the classical theory: (1) Say's law, (2) the quantity theory of money, and (3) continual clearing of the labor market at "full employment."

Keynes's informal model (*informal* meaning that he did not specify his model with a series of equations or represent it with a set of diagrams but, rather, mainly used verbal exposition) begins by seeing the classical economists as having all held the validity of Say's law of markets. Say's law denoted an argument that all income generated by production would be spent on purchasing the national product—either directly, as when workers' wages or capitalists' profits are spent on consumption, or indirectly, when capitalists' savings are borrowed by firms to finance purchases of new capital goods. In modern economic language, we say that the classical economists assumed that the level of *aggregate demand* (Keynes's innovative term) for products always equals the cost of producing them, including a return on their capital to the capitalists whose firms produce the products; in other words, aggregate demand is equal to aggregate supply. Because this would mean that firms would always be able to sell any quantity of goods they might produce, they would choose output levels based on their calculations of their profit-maximizing outputs. They would collectively find it profitable to employ as many workers as were willing to sell labor services at the going wage rates; full employment would

result: The only unemployed workers would be those unwilling to accept the going wage rate, who were considered to be "voluntarily unemployed." Keynes rejected this proposition; in his model, aggregate effective demand for the social product was quite likely to be less than the value of what would be produced if all workers willing to work were employed (aggregate demand could be less than aggregate supply at full employment). If the latter occurred, firms would reduce their employment and output levels until they arrived at the levels at which their sales rates equaled their production rates. So aggregate demand determines the level of actual output and employment, and significant and persistent involuntary unemployment can occur. The cause of recessions or depressions is inadequate aggregate demand. How could this occur? Keynes had a simple answer: Although workers' wage income was generally spent, creating demand for products, profits and interest might very well be saved. And if all of those savings were not borrowed to finance demand for investment goods, aggregate demand could be less than sufficient to employ all those willing to work. Depressions are caused by "too much saving"—too much in comparison with the amount of investment that firms are willing to borrow to finance

Keynes also rejected the validity of another aspect of classical economics: the quantity theory of money (QTM). The QTM argues that the exchange value of a monetary unit is expressed in the average level of prices within the economy (P), and that increases or decreases in the quantity of money (M) in circulation would cause changes in the price level P. Sophisticated versions of the theory were presented using the equation of exchange $M \times V = P \times Y$, in which V represents the velocity of circulation of the money stock and Y represents the current (and assumed full-employment level of national product). If V and Y are assumed to be constant, changes in M would lead to directly proportional changes in P. Because the QTM assumed that V and Y were fixed in the short run, changes in the quantity of money (e.g., an increase in M) would not affect the real level of output or employment. So money is "neutral" within the economy: Changes in the money stock will not cause changes in the levels of output or employment.

In Keynes's model, changes in the quantity of money could cause changes in interest rate levels that could influence aggregate demand for products (especially for investment spending on new capital goods) and thus could influence aggregate production and employment. Money is not generally neutral in Keynes's model; increases in the money stock would affect only the price level and not output in the special case of the economy already being at the full-employment, full-output level.

Finally, Keynes also rejected the validity of the classical views on the working of the labor market, using Pigou's model as an example of these views. Pigou and the other neoclassical marginalist theoreticians had argued that if wages were flexible because of competition for employment and employees, the levels of both the money and real wage (the purchasing power of the level of nominal or money wages) would adjust to "clear the market" for labor. They also argued that the equilibrium *market-clearing real wage* would be equal to the marginal product of labor. Keynes argued that money wages are usually fixed in the short run and that workers cannot bargain over their real wage that would clear the market, because neither they nor their employers can set output prices while bargaining over money wage levels.

## Keynes's Model Represented as a Set of Propositions

Keynes's theory and model attempt to describe the relationship between the amount of money in circulation, the level of interest rates, and the level of employment (his model, the GT, places financial markets and interest rates at the heart of the macroeconomy).

1. High unemployment is caused by insufficient aggregate demand for national product.
2. Insufficient aggregate demand is the result of saving in excess of business' willingness to borrow to finance investment in new capital goods.
3. Excessive saving is done by high-income earning households, who receive interest on bonds; it is the result of an "arbitrary and excessive inequality in the distribution of income" (Keynes, 1936/1973, p. 372). Low investment is caused by low profit expectations or high interest rates, or both—both of which discourage business investment in new capital goods.
4. Low levels of aggregate demand by the private sector (firms and households) can be offset by high levels of government expenditures (stimulative fiscal policy). Low levels of interest rates can also encourage more investment.
5. He argues that the rate of interest should be seen as the price of liquidity (or the price of holding financial wealth as money, as opposed to less liquid financial assets, such as bonds). He termed this the *liquidity preference* theory of money and interest.
6. The level of the rate of interest is explained as determined by the interaction of the supply and demand for money, with the supply controlled by the central bank through monetary policy and the demand determined by the level of transactions and the price level of those transactions, plus financial wealth holders' expectations regarding the future market price of bonds.
7. Another theoretical proposition advanced by Keynes was that if either government expenditures or private investment expenditures increased, total expenditures (aggregated demand) and actual output would increase by a greater amount: He termed this the *multiplier effect.*

Soon after the publication of his book, professional economists began discussing Keynes's views in professional journals, most often in the form of book reviews.

Other economists attempted to describe and interpret Keynes with more formal models. The most influential example of this is John R. Hicks's (1937) IS/LM model, which represented Hicks's interpretation of Keynes's theory with a set of equations and a two-dimensional diagram. Hicks argued that Keynes's theory rested on the validity of his assumption that money wages were fixed in the short run and would not fall if unemployment increased, so Keynes's model should be seen as valid only in the "special case" of an economy in which some institutional factors prevented wages from being downwardly flexible; ironically, Keynes had not presented a "general theory," valid for all economies. Hicks later repudiated his evaluation of Keynes's theory and argued that his model oversimplified Keynes's theory (Hicks, 1957, 1967).

Other attempts at presenting a formal model of Keynes's theory were made by Alvin Hansen (1953) and Paul Samuelson (1948). Their model became known as the Hansen-Samuelson *Keynesian cross diagram.* Given the levels of interest rates, money wages and prices, the parameters of the consumption function—that is, the relationship between personal income and personal consumption expenditures—and planned business investment expenditures, the economy finds a unique equilibrium level of output at the level where aggregated demand equals aggregate production.

## Keynesian Economics

By the early 1950s, Keynes's theory was widely seen as valid by most professional economists (especially among academics). Economists accepting the validity of the Keynesian short-run theory of output determination became known as *Keynesians.* Their published work focused on how fiscal policy could be used to prevent or end depressions, but they also developed business cycle theory and growth theory consistent with Keynes's short-run model. Toward the end of the 1950s and in the early 1960s, some of these economists began to argue that Keynes's theory of depression (inadequate demand causing involuntary unemployment) could be extended to explain price inflation (too much demand at full employment). Logically, there should be a level of unemployment consistent with stable prices and wages, and this was seen as a desirable target for macroeconomic policy. *Full employment* became defined as the lowest rate of unemployment consistent with wage and price stability. Empirical work attempting to estimate the relationship between wages, prices, and unemployment levels produced evidence of an inverse relationship between unemployment and wage or price inflation, consistent with theory. Because such a relationship is downward sloping if graphed with inflation on the vertical axis and unemployment on the horizontal axis, this curve became a part of the core of Keynesian economics (and named the *Phillips curve* after A. W. Phillips, 1958, who conducted one of the earliest econometric estimates of its parameters). Another core belief among the "original Keynesians" was that monetary policy by itself was not usually sufficient to end a severe recession; rather, a fiscal policy stimulus of higher spending by the national government was required. The increases in U.S. gross domestic product (GDP) following increases in expenditure during the 1930s and especially the dramatic increase following large increases in World War II (WWII)–related military expenditures was interpreted by the Keynesians as empirical evidence in support of their central propositions.

However, even as Keynesian economics became dominant in the academy and widely adopted as a guide to macroeconomic policy, widespread opposition to the core theory appeared within academic economics, and Keynesian economics itself divided into several quite different schools of thought, stressing different aspects of Keynes's somewhat vague and informal model in the GT, including approaches now known as *original Keynesians*, *post-Keynesians*, and *new Keynesians*.

## The Evolution of Macroeconomics Since the 1960s

The history of macroeconomics over the last 40 years can be interpreted as a struggle between competing visions of the economy and the proper macroeconomic roles of the state. One vision is exemplified by *new classical economics,* which sees the economy as essentially stable and tending toward an equilibrium characterized by high employment and an economic growth rate largely determined by the rate of technological change (the *natural rate of unemployment* and the *steady-state* rate of growth). Another, contrasting, approach taken by institutionalist and post-Keynesian economics rests on a vision of a very unstable economy, whose growth rate is the result of an open-ended transformational process taking place through economic fluctuations, characterized by excessive unemployment and inequality, and which is often threatened by incoherence and the possibility of breakdown; this approach is called *evolutionary Keynesianism* here.

The first approach implies a noninterventionist role for the state in the economy; the second argues for a strong interventionist state. A third approach, which acknowledges occasional episodes of instability and a limited role for the state in stabilization and the active promotion of growth, appears in new Keynesian economics and new endogenous growth theory, and appears currently hegemonic within mainstream economics.

New classical economics supports arguments against activist macroeconomic stabilization policy, whereas original Keynesian, new Keynesian, institutionalist, and post-Keynesian economics all support intervention. To a great extent, the debates within macroeconomics reflect a

broader and deeper philosophical division between proponents of a radical laissez-faire economic philosophy and those who advocate a strong, interventionist state with wide responsibilities for the common good. The institutionalists and post-Keynesians share a common vision and models that are similar in many respects; their approach is designated *evolutionary Keynesianism* in what follows. The primary purpose of the rest of this chapter is to delineate the important differences between the new classical (NC), new Keynesian (NK), and evolutionary Keynesian (EK) approaches. This chapter suggests that although new classical economics (NC) dominated the mainstream in the 1970s and 1980s, its influence has declined recently as new Keynesian (NK) economics has become more influential, dominating the postmonetarist "new consensus" among macroeconomists. The chapter also argues that the institutionalist/post-Keynesian approach offers a third framework for economic policy.

## The Rise of New Classical Economics: Monetarism, Rational Expectations, and Real Business Cycle Theory

The brief hegemony of Keynesian economics within academia and policy economics lasted for about 2 decades (perhaps from the late 1940s to around 1970) before being challenged by the new version of classical economics. The 1960s consensus approach to macroeconomics was represented with fixed-price, fixed-wage versions of Samuelson's (1948) and Hansen's (1953) Keynesian cross and John Hicks's (1937) IS/LM macroeconomic models of income and employment, and was based on the assumption of a stable relationship between unemployment and price or wage inflation, represented with a Phillips curve. This approach was often referred to by Paul Samuelson's term *neoclassical-Keynesian synthesis,* because its implicit microeconomic foundation was largely a form of the Marshallian price theory that Keynes had used in his GT and that then formed the core of standard neoclassical academic microeconomics, while its macroeconomics was an interpretation of Keynes's theory of effective demand.

The first generation of Keynesian economists saw national governments as responsible for economic stability, economic growth, and full employment, with the highest short-term priority to be given to full employment. Business cycles were seen as caused by fluctuations in aggregate demand, which could be offset with fiscal policy; monetary policy should be used to support fiscal policy, but monetary policy by itself was generally seen as inadequate to stabilize the economy.

Many economists and historians see Keynesian economics as one of the cornerstones of the post–WWII consensus regarding the proper relationship between the state and the economy, and as an integral component of the argument for a more interventionist role of the state. The new political-economic system that emerged in the 1940s and 1950s in most of the wealthy capitalist nations has been described with many terms including *welfare capitalism, managed capitalism, state capitalism, monopoly capitalism,* and *guided capitalism.* The counterrevolution to the Keynesian revolution of the 1940s and 1950s began in the late 1960s with the rise of monetarism, the first component of what was to become NC economics.

### Monetarism

Monetarism can be understood as a set of theoretical propositions and policy proposals focused on the macroeconomic role of money. The core theoretical propositions are as follows:

1. The economy is essentially stable and tends toward an equilibrium at a "natural rate" of unemployment that is consistent with stable wages and prices; this natural rate of employment, the stock of capital, and technology then determine the potential level of national income.

2. Disturbances to equilibrium are almost always caused by changes in the money stock or its growth rate.

3. If unanticipated, these monetary shocks can result in temporary fluctuations in output and employment, but the economy tends to return to the natural rate quickly as wages, prices, and interest rates adjust. Monetarists argue that anticipated monetary shocks are likely to change only the levels of wages, prices, and interest rates even in the short run. Changes in the money supply are neutral in the long run with respect to the real dimensions of the economy: the levels of output, employment, the composition of output, and the real wage. This proposition is described as the *classical dichotomy* between the real and nominal or monetary dimensions of the economy, and its refutation was one of the core ideas in Keynes's GT.

4. Inflation and deflation are the result of excessive or insufficient growth rates in the money stock as in the QTM.

A set of ancillary propositions agreed to by most monetarists was consistent with and supported the core distinguishing theoretical principles of their school mentioned previously.

5. Most versions of monetarism assumed that fiscal policy could not be used to stabilize an economy or otherwise improve macroeconomic performance; this was described as the "ineffectiveness" of fiscal policy in the literature. Changes in government budgets (deficits) intended to stimulate the economy that were financed by borrowing and bond issue resulted in rising interest rates and "crowded out" private investment. Deficits financed by printing money led to inflation; expenditures financed by taxes lowered private spending commensurately. In any of these scenarios, output composition and the allocation of resources would change, but not the aggregate level of

output or employment; surpluses reduced interest rates and stimulated private investment. Because the private sector usually used resources more efficiently than the state sector, expansionary fiscal policy would have detrimental effects; because the economy tended toward the natural rate of employment, countercyclical fiscal policy was unnecessary as well as ineffective.

6. The monetarists argue that the normal state of the economy is the natural rate of employment and positive economic growth as described in Robert Solow's (1956) neoclassical growth theory. A higher saving rate would lead to a higher capital–labor ratio and raise per capita income during a transitional period, but diminishing marginal returns to capital and the free flow of capital and technology across nations implied that growth rates should converge to a "natural rate" of growth determined by technological change.

The key policy proposal advanced by monetarists became known as *Friedman's rule*: Central banks should concentrate on keeping the price level constant (or inflation very low) by setting the growth rate of money at the anticipated growth rate in real output plus the estimated change in velocity. In more sophisticated versions, Milton Friedman (1968) acknowledged that some circumstances such as financial crises might warrant a temporary abandonment of monetary growth targets by the central bank, but he argued that such episodes would not occur very often if the central bank were committed to a stable monetary growth regime, which would be consistent with a stable economy. If the chief source of fluctuations in nominal GDP were fluctuations in the money supply, economic fluctuations would largely disappear under such a regime.

Monetarists also advocated flexible exchange rate systems, arguing that they would strengthen the effectiveness of monetary policy and increase its independence by doing away with the necessity to use monetary policy to peg the exchange rate. Keynes and the Keynesians favored interest rate targets, discretionary monetary policy as a supplement to fiscal policy, and fixed exchange rates to reduce uncertainty.

As increasingly formal versions of monetarism were presented, debates centered on how expectations regarding future levels of wages and prices were formed, and how changes in expectations affected the relationship between unemployment, wages, and prices, which the first generation of Keynesians had thought to be fairly stable and described with Phillips curves. Early versions of Phillips curves (Phillips, 1958; Samuelson & Solow, 1960) described a stable inverse relationship between wage or price inflation and unemployment, which would allow policy makers to choose a level of unemployment and inflation. Edmund Phelps (1968) and Friedman (1968) argued that Phillips curves shift over time, implying unanticipated changes in the natural rate, as the economic environment changes; especially important in their view were expectations of future inflation, which were largely determined by the recent past behavior of prices.

If monetary policy caused inflation by pushing unemployment below the natural rate (by "surprising" workers who did not anticipate reductions in their real wage caused by the inflation), the increase in inflation would lead to more inflation as economic actors attempted to regain their real income by raising wages, prices, and interest rates (known as the *Gibson paradox* and *Fisher effect* in the literature). Economic actors' inflationary expectations adapted to the actual rate of inflation as they looked backward into time trying to forecast economic conditions. These related propositions came to be known as the backward-looking or *adaptive expectations* model and were represented with inflation-augmented Phillips curves, which were vertical at the natural rate in the long run, as in the "neutrality of money" story.

By the late 1970s, monetarism, the natural rate hypothesis, shifting Phillips curves, and vertical long-run Phillips curves at the natural rate of unemployment appeared in all macroeconomics textbooks and were widely accepted as valid analytic concepts within mainstream economics. Support for the neoclassical-Keynesian synthesis and discretionary countercyclical fiscal policy declined. The Federal Reserve began targeting monetary aggregates in 1970 and increasingly emphasized control over monetary aggregates as its primary intermediate target for policy and low inflation as its primary ultimate objective throughout the 1970s and early 1980s; between 1979 and early 1982, it conducted an inflation-fighting monetarist "experiment," targeting the growth rate of the monetary base and the monetary aggregates M1 and M2, while allowing interest rates to increase and fluctuate widely. This episode was consistent with the wide support for monetarism within economics and is often cited to indicate the high watermark of monetarist influence among policy makers in the United States.

## Rational Expectations

A parallel development beginning in the early 1970s was the increasing insistence by some economists that Keynesian economics was not based on the proper *microfoundations* with respect to assumptions about human behavior. If economic actors are rational and utility maximizing, and markets are complete and efficient, markets should continuously clear—including the market for labor.

Persistent involuntary unemployment seems logically inconsistent with those assumptions, since the labor market should allow utility-maximizing workers and profit-maximizing firms to find each other. Many economists began to reject Keynesian models as unscientific because they ignored these issues and seemed inconsistent with the rational expectations hypothesis.

Combining aspects of monetarism (the quantity theory) and the Walrasian microfoundations critique, Robert E.

Lucas Jr. (1972, 1973, 1975, 1976), Thomas Sargent and Neil Wallace (1975), and others argued that if changes in the money stock caused inflation, and if economic actors understood the connection between money and prices—and if they were rational—they would come to anticipate inflation whenever the money stock grew (they would learn from their mistakes and change their behavior). If rational economic actors noticed that the central bank increased the money supply whenever unemployment increased, their "rational" reaction to increasing unemployment and anticipated money supply growth would be to raise wages, prices, and interest rates. Rational economic actors would be forward looking in forming their expectations regarding inflation. If so, monetary policy could not be effective in changing the real dimensions of the economy in even the short run unless the policy changes were unsystematic—irrational policy moves that rational actors would not anticipate, such as raising interest rates in a recession or lowering them in an inflationary boom.

Building on the rational expectations framework, Robert Barro (1974, 1981a, 1981b) argued that rational behavior by forward-looking economic actors would also prevent fiscal policy from stimulating the economy. For example, if the government proposes a tax cut to stimulate consumption and employment, rational consumers and tax payers will anticipate higher taxes (on themselves or their descendants) in the future to repay the increased government debt and will save more to finance those higher anticipated taxes, reducing their current consumption: Aggregate current demand cannot be stimulated with tax cuts. This proposition is known as *Ricardian equivalence* because Barro claims David Ricardo as an early proponent (although Ricardo himself seems not to have believed in the empirical validity of the proposition; O'Driscoll, 1977). The rational expectations hypothesis thus supports arguments for the irrelevance of both fiscal and monetary stabilization policy. Note that the critical assumptions supporting the ineffectiveness of intervention is that changes in the money supply will lead always to changes in the price level and not cause changes in the real dimensions of the economy (the monetarist quantity theory and neutrality of money hypotheses), and on a more fundamental level, that the economy is stable and tends toward the natural rate of unemployment, which is derived from the market-clearing hypothesis. Rational expectations economists often describe their models as *equilibrium economics.*

The rational expectations theory began to dominate economics as taught in elite graduate programs in the late 1970s, appearing in textbooks at about the same time. By the early 1980s, its radical argument against the possibility of altering the real dimensions of the economy through monetary or fiscal policy was widely accepted within the profession and became dominant by the end of the decade.

In the late 1970s, versions of the NC theories, business cycles, and inflationary episodes were mainly the result of external shocks to the economy, temporary misperceptions by workers or firms regarding the wages or prices that would clear markets (false trading), or the unintended consequences of well-intentioned but doomed attempts to stabilize the economy, and the latter were most often caused by unanticipated attempts to push the unemployment rate below the natural rate and to raise growth above the long-run trend determined by growth in resources and technological change. The resulting inflation required central banks to tighten monetary policy, forcing the economy into a recession until inflationary expectations were reduced and the economy returned to its equilibrium natural rates of unemployment and growth. The state should leave the economy alone, as in most versions of original classical economics; economic policy should be restricted to providing the proper institutional framework for a capitalist market economy and instructing the central bank to follow Friedman's rule. This view is a profound rejection of the political economy of Keynes and the early Keynesians, who held that the state can and must improve the performance of the economy through discretionary monetary and fiscal policies.

## Real Business Cycle Theory

But to many of those working within the rational expectations–Walrasian model, the "bad monetary policy" story seemed an inadequate explanation for business cycles; real business cycle (RBC) theory emerged in the early 1980s to offer an explanation consistent with both the Walrasian continuous market-clearing approach and the rational expectations theory's definition of rational behavior. In this theory, economic fluctuations are largely the result of "real" or nonmonetary factors: changes in technology and in preferences by workers for leisure versus goods, or *intertemporal substitution.* Expansions and high growth rate periods are the result of the introduction of new technology sets, as in Schumpeter's theory of economic development (and Marx's as well); recessions occur when the economy readjusts to the diminishing influence of a set of technological changes on investment. Increases in unemployment not caused by technological change are the result of workers' preferences shifting away from products toward leisure, or rational responses to changes in real wages and interest rates that alter the relative prices of goods and leisure.

Very importantly for present purposes, the RBC models denied any real effects of changes in the money stock on the economy; in fact, in an interesting twist, the money supply was seen as endogenous to the economy: The money stock increased in expansions and declined in contractions, passively reacting to cyclical changes in the demand for loans (as in the endogenous money supply theory advanced by the EK school). The RBC theory is based on a radical version of the classical dichotomy between the real and monetary dimensions of the economy.

The emerging new version of classical theory that attacked the previous Keynesian consensus began with an argument that only changes in the money stock could influence the economy (early monetarism); moved to the position that money mattered but only in the short run (the inflation-enhanced Phillips curve version of monetarism); then adopted the rational expectations position that only unanticipated, unsystematic, irrational monetary policy ("bad policy") could have even short-run effects; and then finally proposed that money did not matter at all with respect to causation in the short-run or long-run behavior of the economy in the RBC models.

The NC theory presented a view of the economy consistent with the original classical story at least in its popularized form within modern economics literature: Capitalism is self-adjusting and stable; competitive markets lead to the most desirable state of affairs; the normal state is high employment with economic growth at the highest rate possible, given time and leisure-goods preferences and the exogenously determined rate of technological change. The only policy role for the state consistent with this vision is providing the necessary institutional structure; otherwise, laissez-faire and free market fundamentalism are advised. By the late 1980s, NC dominated academic economics in the United States in the elite graduate programs and in textbooks, and was widely taught to undergraduates as well.

## Opposition to the New Classical Theory

Two strong currents questioning the validity and challenging the hegemony of NC macroeconomics developed even as NC emerged: the new Keynesians and the institutionalist/post-Keynesians or evolutionary Keynesians (EK). The new Keynesians operate within the mainstream, teaching at elite universities and publishing in the profession's highly ranked journals; many of them (e.g., Ben Bernanke, Alan Blinder, Stanley Fischer, Gregory Mankiw, Joseph Stiglitz, Lawrence Summers, John Taylor, and Janet Yellen) have held important policymaking positions in institutions such as the Federal Reserve, World Bank, Council of Economic Advisors, U.S. Treasury, and the International Monetary Fund.

Meanwhile, outside of the inner circle of mainstream economics, a radical critique of both NC and NK based on a different vision of the economy, a different set of assumptions about human behavior and economic reality, and perhaps a different set of social values and priorities was developed by the EK school.

### New Keynesian Macroeconomics

NKE emerged in the late 1970s and early 1980s in reaction to the criticism of consensus IS/LM Keynesianism mounted by the NC school; their emphasis was on explaining the causes of business cycles. The new Keynesians retained Keynes's insistence that economic fluctuations can be caused by aggregate demand changes and that aggregate demand fluctuations could be caused by factors other than monetary shocks, and they retained the early consensus Keynesian approach that held that wages and prices were downwardly inflexible in the short run and that recessions could be seen as the result of "coordination failures" and "quantity adjustments" to demand or supply shocks. They retained the Keynesian view that recessions were inherent in capitalism, undesirable, socially expensive, and preventable with correct policy. But many of them accepted the NC microfoundations argument that assuming rational utility-maximizing behavior by economic actors and some version of rational expectations was necessary and useful for economic analysis.

Although some NK economists have advocated the use of countercyclical fiscal policy in severe recessions or when the threat of deflation appears (Stiglitz, 2002; comments by Auerbach, Blinder, and Feldstein in Federal Reserve Bank of Kansas City, 2002), most of them have argued that monetary policy is more efficient and generally sufficient to stabilize the economy. And although they advocate interventionist monetary policy to stabilize the economy (money is not neutral in the short run), they generally express a preference for monetary policy rules as opposed to discretion (Taylor, 1999, 2000). "Rules" means setting targets for policy (the rate of inflation, or more often minimizing the gap between actual and *potential GDP,* defined as the level of GDP consistent with the lowest sustainable level of unemployment without accelerating price inflation—the nonaccelerating rate of inflation [NAIRU]) and designing a policy reaction function in which the central bank would increase or decrease interest rates by a given amount if GDP exceeds or falls below potential. Most of the new Keynesians see the money supply as endogenous in the sense of its growth rate being the interaction of demand for credit and the central bank-determined level of short-term interest rates, and see money as neutral in the long run with respect to its influence over the growth path of the economy (DeLong, 2000; Parkin, 2000; Taylor, 2000).

Their primary focus and contribution with respect to understanding business cycles has been to demonstrate that (a) inflexible wages and prices could lead to quantity adjustments that were destabilizing (recessions could be understood as systemic coordination failures of the economy's markets because markets do not always quickly find prices that "clear") and that (b) rigid, sticky, or slowly adjusting prices and wages could be seen as the result of rational responses by economic actors within the actual institutions of capitalist economies. Much of their research focused on the latter point; they found plausible explanations for sticky wages and prices, which challenged both the NC argument that markets clear quickly (or would in a more nearly perfect world) and the NC claim that the Walrasian equilibrium approach was useful as a description of reality. NK has been described as *dis-equlibrium*

economics, in that it explains why economies are usually not in equilibrium at the natural rate of unemployment and the potential level of output.

Inflexible wages and prices are explained by institutional factors such as monopolistic competition, menu costs, lengthy contracts, efficiency wage theory, wage and price staggering, markup pricing, bureaucratic inertia, and marketing strategy. Other NK lines of attack on NC involved skepticism regarding aspects of rational expectations, importantly including the assumption of inexpensive and complete information; the NC assumption that workers, managers, and owners actually think and behave like the NC economists' models would have them; and propositions built on rational expectations, such as the Ricardian equivalence story.

In summation, the new Keynesians argue that capitalism is often unstable due to the persistence of both demand and supply shocks, and to the ways in which the market system adjusts to such shocks. They also believe that it is both desirable and necessary to use interventionist policy to stabilize the economy; these positions put them in the Keynesian camp and in clear opposition to the views of the NC school. Their preference for monetary policy over fiscal policy—(a) because they think monetary policy is usually effective, (b) because they think that automatic stabilizers are more effective than discretionary policy due to the relatively small estimates for multipliers and the long time lags for fiscal policy, and (c) because of the political problems that make timely changes in fiscal policy difficult—and for policy rules over discretion differentiates them from the original Keynesians as well as the post-Keynesians.

The views of the new Keynesians appear to be currently hegemonic within mainstream academic macroeconomics in the United States and United Kingdom from the perspectives of who has been selected for policy advice by the Federal Reserve, the Bank of England, and recent governments in both countries; whose macroeconomics texts are most widely adopted and whose theoretical views are dominant in classrooms; and whose policy views are adhered to by the Federal Reserve.

### New Economic Growth Theory

Another interesting aspect of mainstream economics is the development of new neoclassical approaches to economic growth that go beyond the Solow growth models and lend support to arguments for government intervention in the economy. Solovian growth models are based on neoclassical and NC assumptions such as perfect competition (and continuous market clearing), diminishing marginal returns to capital, the free flow of information and technological change, and equilibrium between aggregate demand and aggregate supply (so that the economy is assumed to be always at full employment). In these models, diminishing returns to capital lead to the counterintuitive deduction that high rates of investment will have no effect on economic growth over the long run; *conditional convergence* should be obtained, in which countries with similar savings and population growth rates should converge to the same level of per capita national income and to the same rate of growth (the stationary state), while countries with different characteristics would end up with different per capita income levels but the same growth rate. The common growth rate would be determined by the exogenous rate of technological change under the assumption that technology and knowledge are mobile across countries (Solow, 1956); this leaves almost no role for the state in promoting economic growth.

The new economic growth theories (NEG) broaden the definition of *capital* to include knowledge (human capital) and also incorporate the spillover effects of investment in both human and fixed capital and the effects of increasing returns to scale. Under these conditions, countries with higher rates of savings and investment could have permanently higher rates of technological progress and economic growth. Because the growth rates of technological progress and output are influenced by the rate of investment, which is determined within these models, this approach is often termed *endogenous growth theory.*

For present purposes, the importance of the NEG models is that they present another reason for state intervention in the economy: The state can encourage economic growth (a) through its investment in human capital (education, research, and development) and in infrastructure (Aschauer, 1989), (b) by developing appropriate institutions (competitive markets, well-regulated financial systems, stable money), and (c) through policies that encourage saving and investment by the private sector.

## Post-Keynesian Economics

A radical critique of original IS/LM Keynesianism, NCE—especially its monetarist core—and new Keynesianism as well has been developed by institutionalist and post-Keynesian economists, whose separate views on macroeconomics have been merging since the late 1970s. Following the much admired Joan Robinson, the first generation of economists who referred to themselves as post-Keynesians such as Paul Davidson (1978) used derogatory terms such as *bastard Keynesianism, IS/LM Keynesianism,* and *textbook Keynesianism* to refer to what they saw as a much attenuated and misleading version of the master's views that had been developed in the 1940s and 1950s by the first generation of Keynesians. They argued that IS/LM Keynesianism ignores Keynes's stress on uncertainty and disequilibrium; it is another form of general equilibrium theory, describing a tendency toward equilibrium that does not exist in the real world (or in Keynes's theory).

These economists attempted to go beyond Keynes's work (*post*-Keynesian) by building on what they saw

as Keynes's correct ideas and insights while rejecting what they found inadequate in his work. Included in the latter category are his GT assumption of a central bank–determined exogenous money supply (although Keynes apparently held to an endogenous theory of money in his other works), his Marshallian price theory, and the lack of a theory of economic growth. The post-Keynesian project is to construct a realistic model of modern capitalism that would be useful in designing policy to encourage full employment, stability, growth, and less inequality. A strong emphasis on finding practical solutions to economic policy problems is found throughout the work of this group, which has also been a hallmark of American institutionalism. Many of the early post-Keynesians were (and some still are) sympathetic to some versions of democratic socialism; most advocate some form of incomes policy; and all advocate a powerful, interventionist state whose economic policies should give highest priority to encouraging full employment, economic growth, and less inequality—the goals proposed by Keynes in the final chapter of his GT.

The core propositions of post-Keynesian economics (the first two form their "pre-analytic vision," in Schumpeter's [1954, pp. 41–42] term) include the following:

1. The recognition of *fundamental* or *absolute uncertainty* as radically different from *statistical* or *probabilistic risk;* fundamental uncertainly does not allow us to make precise calculations of risk. Keynes observed that many of the most important economic decisions—such as whether to invest in fixed capital, purchase a bond, or hold money—are made in situations in which the information necessary to evaluate risk probabilistically will always be absent. The post-Keynesians' understanding of uncertainty is related to their stress on the importance of historical time and the irreversibility of many important decisions, and is antithetical to the NC approach to knowledge and uncertainly.

2. The economy is inherently unstable because of uncertainty and the instability of expectations, especially expectations regarding profit from investment and the future prices of assets. The classical and NC's equilibrating mechanism of flexible prices is weak (prices are not very flexible downward), and it would actually increase instability if prices could somehow be made more flexible with institutional change, because falling prices in a recession would depress profit expectations and investment. Full employment is less likely than widespread unemployment. Financial speculation and financial instability are inherent in the structure of modern financial institutions and financial markets, and can be the cause of instability in the "real" economy. Society needs to impose stabilizing constraints on the economy, including institutions that stabilize prices, wages, and interest rates. Most important, a large government sector whose expenditure can quickly increase in slumps is necessary to prevent downward instability (Minsky, 1982); small state sectors reduce stability.

3. Economic growth, economic fluctuations, and income distribution are dialectically related and mutually reinforcing: Inequality enhances instability; instability (especially recessions) reduces investment and growth, while recessions and low growth increase inequality. This centrality of demand in post-Keynesian theory leads to the proposition of *demand-led growth:* The long-run growth path of the economy is determined by its short-run behavior.

4. Economies are best understood as "complex systems" that are "self-organizing" and exhibit "emerging properties" (see Moore, 1999, for a clear statement of this proposition and its implications; it is related to the institutionalist economists' insistence that society's institutions evolve through time, so that theory must be institutionally specific to be useful).

5. Moving to another core proposition on a lower level of abstraction, the entry point into macroeconomics for the EK school is a "monetary theory of production" (Keynes, 1936/1973). Money is created (by banks) to finance an increase in production, which requires more fixed and circulating capital. Money is necessary for production to take place because production takes time and because money is the social institution that transfers and stores purchasing power over time. Because of the existence of money, interruptions in the circular flow of income and expenditure can take place (by holding wealth in the form of money), which are unlikely in a barter economy.

According to the EK school, the NC's *axiom of reals* (also held to a lesser extent by the NK school as well)—the dichotomy between the monetary and real dimensions of the economy—is misleading; capitalism must be understood as a monetary economy, in which the circuit of money is as important as the physical flow of production and circulation of goods and services. Macroeconomics must begin with an analysis of money—its nature, origin, and functions; money is never neutral with respect to the real economy. The level of interest rates (the price of liquidity and the cost of credit) is a key price within the economy, because it influences both the willingness and the ability of entrepreneurs to invest in real capital and of financial capitalists to hold nonmonetary financial assets such as stocks and bonds.

6. The levels of prices, wages, and interest rates should be understood as the result of distributional struggles that are determined by social institutions and complex processes, as opposed to their determination by the quantity theory of money as advanced by NC. Rather than determining the level of wages and prices, the quantity of money in circulation is seen as the result of changes in the level of wages and prices, reversing the quantity theory's direction of causality. In EK economics, changes in the

level of wages lead to a change in the prices of goods, because firms practice markup pricing: Prices are marked up over—primarily—labor costs of production, and estimates of average rather than marginal cost at normal output levels are used to set prices. Changes in the price level (inflation) leads ceteris paribus to an increase in the demand for working capital (credit), which banks accommodate. As more loans are made, the money supply increases (Moore, 1988).

7. The endogenous money supply theory argues that anything that increases the demand for bank credit will increase the money supply; commercial banks must accommodate most of any increase in business or consumer loan demand, because most loans are made under predetermined lines of credit (Moore, 1988).

8. Finally, although EK economists see the level of interest rates as important in influencing aggregate demand—especially business investment—their empirical work argues that spending and real output may not be very sensitive to interest rates in recessions, so that other demand-stimulating policies are necessary (Arestis & Sawyer, 2002a, 2002b; European Central Bank, 2002). Those same studies provide evidence that interest rate changes are not very effective in reducing price inflation either.

## Post-Keynesians and New Keynesians

Although the EK and NK economists agree on aspects of macroeconomics (e.g., the endogeneity of the money supply and the need for interventionist demand management), they disagree on many important points.

1. EK economists see the economy as very unstable, requiring constraints stronger than discretionary monetary policy; new Keynesians assume strong equilibrating processes in the long run, pushing the economy toward equilibrium at a socially optimal natural rate of employment and the potential level of output.

2. EK economics follows Keynes (and Kalecki) in arguing that savings does not finance investment as in the old classical, neoclassical, NC, and NK models. Rather, investment is determined by profit expectations and the rate of interest, and it is financed by bank credit (and the growth in the money supply). The level of investment coupled with the variables that determine the Keynesian multiplier then determines the level of national income. National income moves toward the level that generates enough savings to equal the exogenously determined level of investment: Investment determines savings, rather than savings determining investment.

This insight has powerful implications for many aspects of policy: Most NC and NK economists argue for policies (such as low marginal rates of taxation, high real interest rates, shrinking government, reducing the generosity of public pension systems) that should encourage higher net national savings (savings net of government budget deficits and depreciation), because in their models, this would lead to higher private investment. From the EK perspective, this is wrongheaded: Government spending—especially public investment—can increase productivity, private profits, and profit expectations, thus encouraging private investment ("crowding in" rather than "crowding out"). And the level of savings has little influence over either interest rates or investment, because interest rates are primarily determined by monetary policy and liquidity preference. In fact, *ceteris paribus,* a higher saving rate might depress investment and economic growth because it would lead to a lower level of consumption and aggregate demand growth.

3. EK economists put a higher priority on full employment than on price stability and argue that the level of unemployment necessary to keep effective downward pressure on wages and prices entails unacceptable social costs. NK economists put a higher priority on low inflation than on full employment. EK economists are skeptical of our ability to reliably estimate the level of unemployment consistent with price and wage stability—the natural rate of unemployment or the NAIRU, which determines the potential level of national income in the NK model—and use that as a target for stabilization policy. There is no natural rate of unemployment in the sense of a *strong attractor* that the economy tends toward.

EK economists argue that the important goals of full employment and both wage and price stability can be reached only by developing institutions that socially control wages, prices, and the distribution of income across the social classes (the wage–profit ratio), and that link aggregate wage and profit increases to productivity gains.

In contrast, new Keynesians argue that the NAIRU can be reliably estimated (although the range of estimates is seen by some as quite wide) and the estimates used as a target for stabilization policy; NK economists are willing to tolerate whatever levels of unemployment are necessary for price stability, disagreeing with the EK view that "non-traditional" forms of intervention such as incomes policy can be effective.

4. EK economists agree with both Keynes and Kalecki that the distribution of income is important for business cycles and growth; they are interested in both the *functional* or *class* distribution between labor and capital and the *personal* or *size* distribution across individuals, households, and families. One argument that Keynes and post-Keynesians make is that changes in the distribution of income can influence the composition and levels of aggregate demand (more profits means more investment, higher wages means more consumption). From this perspective, less inequality is preferred because it will stimulate production and employment in the short run and

thus stimulate investment and economic growth in the long run; lower interest rates both reduce inequality and stimulate investment (Keynes, 1936/1973; Niggle, 1998, surveys and assesses some recent literature discussing the relationship between inequality and growth).

Followers of Kalecki observe that a declining wage share should be expected to reduce aggregate demand, capacity utilization, and investment, and thus reduce both employment and economic growth. Proper macroeconomic policy implies paying attention to income distribution. Again, new Keynesians do not pay much attention to these issues.

5. EK emphasizes a demand-led approach to growth theory, in contrast to the NC supply-side approach, which new Keynesians and new (endogenous) economic growth theories also stress.

6. Many EK economists advocate fixed exchange rate systems constructed around an international financial institution that could issue liquid financial assets as needed by deficit countries (Davidson, 1994); most new Keynesians, such as Joseph Stiglitz (2002), accept flexible exchange rates with some important and influential exceptions.

7. EK economists favor financial market regulations and see unregulated financial markets as dangerous (Isenberg, 2000; Minsky, 1982, 1986); most new Keynesians are not concerned with financial market deregulation.

## Conclusion

In the 1970s and 1980s, NC economists developed and mainstream economics assimilated a set of propositions, models, and theories that argued against both the need for and the efficiency of Keynesian forms of state intervention in the economy to promote full employment, stability, equality, and economic growth. Aspects of this economic philosophy—NC, monetarism, RBC theory, supply-side economics, and public choice theory—offered theoretical support for neoliberalism and have been very influential both within economics and within the domains of policy and politics.

Keynesian economists rejected many aspects of the neoliberal program based on their competing NK theoretical stance. EK economists present a more radical critique of NC and offer a very different perspective on the economy. In the past decade, new Keynesian and new growth theory economics have become more influential within the mainstream. This phenomenon, coupled with the persistence of economic problems that seem intrinsic to unregulated global capitalism—such as stagnation, increasing unemployment and inequality, and recurrent financial crises—opens up the possibility for EK economics to be seriously considered by mainstream economists, because it offers coherent explanations for those problems along with plausible solutions to them.

## References and Further Readings


Arestis, R., & Sawyer, M. (2002a). The Bank of England macroeconomic model: Its nature and implications. *Journal of Post Keynesian Economics, 24*(4), 529–546.

Arestis, R., & Sawyer, M. (2002b). *Can monetary policy affect the real economy?* (Working Paper No. 355). Annandale-on-Hudson, NY: Levy Economics Institute. Available from http://www.levy.org/download.aspx?file=wp355.pdf&pubid=111

Aschauer, D. (1989). Is public expenditure productive? *Journal of Monetary Economics, 23*(2), 177–200.

Barro, R. (1974). Are government bonds net wealth? *Journal of Political Economy, 82*(6), 1095–1117.

Barro, R. (1981a). Output effects of government purchases. *Journal of Political Economy, 89*(6), 1086–1121.

Barro, R. (1981b). Public debt and taxes. In *Money, expectations and business cycles.* New York: Academic Press.

Davidson, P. (1978). *Money and the real world* (2nd ed.). New York: Macmillan.

Davidson, P. (1994). *Post Keynesian macroeconomic theory.* Aldershot, UK: Edward Elgar.

DeLong, B. (2000). The triumph of monetarism? *Journal of Economic Perspectives, 14*(1), 83–94.

European Central Bank. (2002, October). Recent findings on monetary transmission in the Euro area. *Monthly Bulletin,* pp. 44–55.

Federal Reserve Bank of Kansas City. (2002). *Symposium: Rethinking stabilization policy.* Available from http://www.kansascityfed.org/PUBLICAT/SYMPOS/2002/sym02prg.htm

Friedman, M. (1968). The role of monetary policy. *American Economic Review, 58,* 1–17.

Hansen, A. (1953). *A guide to Keynes.* New York: McGraw-Hill.

Hicks, J. (1937). Mr. Keynes and the classics: A suggested interpretation. *Econometrica, 5*(2), 144–159.

Hicks, J. (1957). The "classics" again. *Economic Journal, 67,* 278–289.

Hicks, J. (1967). Monetary theory and history: An attempt at perspective. In *Critical essays in monetary theory* (pp. 155–173). Oxford, UK: Clarendon Press.

Isenberg, D. (2000). The political economy of financial reform: The origins of the U.S. deregulation of the 1980s and 1990s. In R. Pollin (Ed.), *Capitalism, socialism, and radical political economy* (pp. 245–269). Northampton, MA: Edward Elgar.

Keynes, J. (1973). *The collected writings of John Maynard Keynes: Vol. 7. The general theory of employment, interest and money.* New York: Macmillan. (Original work published 1936)

Lucas, R. E., Jr. (1972). Expectations and the neutrality of money. *Journal of Economic Theory, 4,* 103–124.

Lucas, R. E., Jr. (1973). Some international evidence on output-inflation tradeoffs. *American Economic Review, 63*(3), 326–334.

Lucas, R. E., Jr. (1975). An equilibrium model of the business cycle. *Journal of Political Economy, 83*(6), 1113–1144.

Lucas, R. E., Jr. (1976). Economic policy evaluation: A critique. In K. Brunner & A. Meltzer (Eds.), *The Phillips curve and labor markets* (Carnegie-Rochester Conference Series on Public Policy No. 1, pp. 19–46). New York: Elsevier North-Holland.

Minsky, H. (1982). *Can "it" happen again? Essays on instability and finance.* Armonk, NY: M. E. Sharpe.

Minsky, H. (1986). *Stabilizing an unstable economy.* New Haven, CT: Yale University Press.

Moore, B. (1988). *Horizontalists and verticalists: The macroeconomics of credit money.* New York: Cambridge University Press.

Moore, B. (1999). Economics and complexity. In M. Setterfield (Ed.), *Growth, employment and inflation: Essays in honor of John Cornwall* (pp. 41–66). New York: St. Martins.

Niggle, C. (1998). Equality, democracy, institutions and growth. *Journal of Economic Issues, 32*(3), 523–530.

O'Driscoll, G. (1977). The Ricardian nonequivalence theorem. *Journal of Political Economy, 85*(2), 207–210.

Parkin, M. (2000). The principles of macroeconomics at the millennium. *American Economic Review, 90*(2), 85–89.

Phelps, E. (1968). Money-wage dynamics and labor-market equilibrium. *Journal of Political Economy, 76*(S4), 678–711.

Phillips, A. W. (1958). The relationship between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica, 25,* 283–299.

Samuelson, P. (1948). *Economics.* New York: McGraw-Hill.

Samuelson, P., & Solow, R. (1960). Analytical aspects of anti-inflation policy. *American Economic Review, 50*(2), 177–194.

Sargent, T., & Wallace, N. (1975). Rational expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy, 83*(2), 241–255.

Schumpeter, J. (1954). *History of economic analysis.* New York: Oxford University Press.

Solow, R. (1956). A contribution to the theory of economic growth. *Quarterly Journal of Economics, 70*(1), 65–94.

Stiglitz, J. (2002). *Globalization and its discontents.* New York: W. W. Norton.

Taylor, J. (Ed.). (1999). *Monetary policy rules.* Chicago: University of Chicago Press.

Taylor, J. (2000). Teaching modern macroeconomics at the principles level. *American Economic Review, 90*(2), 90–94.

# 31

# Aggregate Expenditures Model and Equilibrium Output

Michael R. Montgomery
*University of Maine*

*Aggregate expenditures* (AE), the total spending in an economy on final goods and services over a designated time period, is the core demand-side concept in modern macroeconomics (a *final good* is a newly produced good bought by a user who will "finally" dispose of that good by using up its services). Following the lead of John Maynard Keynes in his *General Theory of Employment, Interest and Money* (1936/1965) and early Keynesians such as Alvin Hansen (1953) and Paul A. Samuelson (1939), AE is typically broken down by major type of purchaser into consumption expenditures, investment expenditures, government expenditures, and the sum of exports less imports (known as *net exports*). The study of these categories in recent decades has been aided considerably by the development of the National Income and Product Accounts (NIPA) accounting system, which is used by governments in measuring and reporting the sizes of these categories. The sum of this spending is known as gross domestic product, or GDP. Specifically,

- *Consumption expenditures* are expenditures on final goods (excluding housing) and services by consumers, produced during the accounting period and in the economy under study.
- *Investment expenditures* are final goods and services purchased by businesses and buyers of homes, produced during the accounting period and in the economy under study. They include nonresidential structures of all types, plus all producers' durable equipment, plus residential structures, plus all inventories produced in the accounting period (some of these inventories may be *planned,* or *desired,* inventories; others may be *unplanned,* or *undesired,* inventories—unplanned inventories being acquired when sales are less than anticipated). Investment expenditures do not include financial transactions such as the purchase of stocks and bonds.
- *Government expenditures* are final goods and services purchased or produced by governments at all levels, produced during the accounting period and in the economy under study.
- *Net exports* are exports of final goods and services by the country under study (produced during the accounting period in the economy under study), minus the imports of such goods into that country.

The sum of these four spending components equals AE. AE thus equals GDP. That is, the total value of expenditures on final goods and services is equal to the total value of all that which is expended on the acquisition of final goods and services (i.e., GDP).

## Theory

### Expenditures, Planned Expenditures, and Inventory Adjustment

There are, however, two variants of the AE concept: *actual* AE, as defined previously, and *planned* AE. The latter is of particular importance in modern macroeconomics.

*Planned AE* is the amount of expenditures the economy generates when all purchasers purchase an equilibrium amount of goods and services. It describes a state where purchasing decisions are consistent with the resources and needs of purchasers. (Planned AE often is referred to as *aggregate demand.*)

By way of illustration, consider the behavior of a typical retailer. Each month (let us say), a retailer assesses the state of its business and, based on that assessment, orders a particular quantity of the goods that it is its business to sell. At month's end, any of the ordered goods that are unsold will be added to inventory. If the retailer's assessment of the market demand for its product is correct, its sales of goods will be such that, at month's end, it will be left with just the amount of inventory stocks on its shelves that is consistent with its long-run needs (needs given by relatively stable factors like how many customers the store loses if it runs out of goods its customers wish to buy, the financing costs of carrying such goods, etc.).

Suppose, however, that the retailer has overestimated the demand for its goods, so that it is left at month's end with more unsold goods than it had expected. Had the store known, at the beginning of the month, that demand would be lower, then it would have ordered fewer goods. However, because it did not know this and so did not reduce its orders, some of the goods it had planned to sell to customers that month remain on the store's shelves at month's end. The retailer's actual expenditures on new inventory for the month, then, are greater than its planned expenditures on inventory. The store's actual expenditures on inventory now exceed what it would have spent if it had predicted monthly demand correctly. As a result of its miscalculation, the retailer now has excess, or unplanned, quantities of inventories on its shelves.

As an aid to clarity, consider a simple numerical example. Suppose a T-shirt retailer orders 100 shirts, planning to sell 80 and retain the other 20 as inventory. At month's end, it turns out that the retailer has been able to sell only 70 of the 100 T-shirts that were ordered on the first of the month. It will, therefore, add the remaining unsold 30 T-shirts to its inventory, including 10 more than the 20 shirts it had planned to add. Had the store known that T-shirt demand would drop by 10 units, it would have cut its orders by 10 units, so as not to add more T-shirts than planned to its inventory stocks. Thus, had the retailer been fully informed at the start of the month about the state of demand for its product, it would have ordered only 90 T-shirts.

As things stand, the retailer has ordered 100 T-shirts when, had better information been available at the start of the month, it would have chosen to order only 90. The store's planned investment is the 20 T-shirts it had planned all along to add to its inventories. Its unplanned (or involuntary) investment is the 10 T-shirts it adds to its inventories unintentionally. (Quite likely, next month the store will reduce its orders in order to "work off" its excess stock of inventories, weakening thereby the amount of economic activity at the factories from which it orders its stock of goods.)

The same type of reasoning, but in reverse, holds when the store finds that demand for its T-shirts is greater than expected. Suppose the store has ordered 100 shirts, again expecting to sell 80 and retain 20 as inventory. Suppose demand is greater than expected, equaling 90 T-shirts. The store will then sell all the T-shirts (80) that it had ordered for sale during the month. However, in addition, it will sell 10 more units that it had been planning to add to its stock of T-shirt inventory. The store's planned investment is still 20 T-shirts, but its unplanned investment is –10 (negative 10) T-shirts, due to its unexpectedly strong sales. As a result, at month's end, the store's total investment is 10 T-shirts smaller than had been planned. (Quite likely, next month the store will increase its orders to increase its stock of inventories to its planned levels, increasing thereby the amount of economic activity at the factories from which it orders its stock of goods.)

The simple inventory problem faced by our retailer turns out to be quite important, both as a part of macroeconomic analysis and in the actual economy. The economy often behaves as suggested in the retailer example. In periods of unexpected demand weakness, the economy's retailers, wholesalers, and so on accumulate positive levels of unplanned inventories. Subsequently, retailers and other sellers of goods cut back on orders, causing less production upstream in the factories and so amplifying the initial weakening of the economy. In the reverse case, an unexpectedly strong economy causes an undesired depletion of inventories (negative levels of unplanned inventories), causing increases in orders at the factories and so amplifying the initial strengthening of the economy.

Another reason why the simple inventory problem is particularly important to macroeconomics stems from a characteristic of NIPA accounting. Consider consumption expenditures. How much of the spending that consumers do is truly intended or planned, and how much of it is mistaken (i.e., with full knowledge, particular consumers might have bought more, or less, product than they did in fact buy)? Because of the inherent impossibility of making such a determination, NIPA accounting simply assumes that all of the spending carried out by consumers is planned spending. Similarly, all government spending is considered to be planned spending, and the same convention is applied to net exports. However, for the case of investment spending, as we have already seen, planned expenditures and actual expenditures do differ. In fact, it is in the investment spending accounts that all of the gap between AE and planned AE is found. When consumers, governments, or "net exporters" unexpectedly reduce their expenditures, all of the goods they have chosen not to purchase wind up being held as inventory by merchants who were unable to sell these goods (and vice versa if spenders unexpectedly increase their expenditures). The distinction between planned and unplanned

inventory stocks, therefore, is the mirror image of unexpected changes in expenditures by others in the economy.

## Planned AE and Inventory Adjustment: The Keynesian Cross

The preceding analysis suggests a simple visualization of macroeconomic fluctuations, based on firms' inability to consistently predict demand and the resulting accumulation or reduction of inventory stocks on the firms' aggregate balance sheet. Consider a substantial and unexpected reduction in planned AE in an economy that has grown accustomed to a higher level of such expenditures. As a result, sales fall, and unplanned inventories accumulate on firms' shelves and in their warehouses.

How will retailers, wholesalers, and so on react to the undesired swelling of their inventory? There are two possibilities, in principle. One is for them to immediately lower their prices by enough to induce buyers to quickly clear out their excess inventory by taking advantage of these "fire sales." However, because these firms will likely take a sizeable capital loss on their inventory stocks if they eliminate them in this way, they are often reluctant to lower their prices by enough to sell off very much of their excess inventory stock. The alternative strategy is to hold prices more or less unchanged and to reduce the size of their orders from the factories. This way, firms allow their excess stocks of goods to be worked off slowly over time (because business has slowed, this second strategy usually also involves laying off a portion of the firms' workforce).

To an extent, firms will pursue both of these strategies. However, it is a common observation that firms primarily respond to excessive inventories by laying off workers, cutting orders, and working off inventory stocks slowly over time. This strategy, however, transmits weaker sales "upstream" to wholesalers and factories as described previously, triggering weaker economic activity in these sectors too and causing a lessening of activity and layoffs in those sectors. These layoffs, in turn, reduce demand by consumers in still other sectors where these consumers purchase items, triggering still more unplanned inventory buildup and layoffs in those other sectors, and so on (in a dynamic sequence commonly known as the *multiplier process*). Thus, the initial decline in planned AE that had initially led to firms' accumulation of unplanned inventory investment can, through the preceding sequence, become a general decline in economic activity.

The standard graphical model used to illustrate this process is known as the *Keynesian cross* (after John Maynard Keynes, who founded this line of thought). Sometimes it is instead called the *45-degree line diagram* (after the graph's most notable visual feature). Before discussing this diagram, however, we need to clarify an important relationship in our simple model. One of the core principles of national income accounting is that, allowing for a couple of side issues, the production of a given dollar amount of GDP generates an equal dollar amount of income for the factors of production that participated in the production of that GDP. In fact, this statement is only an approximation of the true state of affairs, due to the side issues of depreciation expenditures and indirect business taxes. In the actual economy, before counting our new production this year, we must replace the equipment we used up (or depreciated) last year due to our productive activities then. Also, in the actual economy, there are indirect business taxes that tax away some business receipts before they can be distributed to the factors of production. However, in our simple illustrative economy, we ignore these two factors. An alternative treatment uses net national product (NNP), which is GDP minus depreciation expenditures, as the core national output concept (indirect business taxes are again set at zero). Using NNP as the output concept is more realistic than using GDP, because depreciation expenditures need not be assumed away as in the discussion in the main text of the chapter.

Returning to the main argument, in our Keynesian cross framework, every dollars worth of production generates exactly a dollar's worth of income to those owning the factors of production. In the aggregate, then, national output (GDP) equals national income by definition. We will take this equivalency seriously, so much so that we will even use the same symbol *Y* to simultaneously represent both national output and national income. That is, in what follows, we will call Y either national output or national income, depending on which better fits the context.

Returning now to our main line of discussion, Figure 31.1 depicts the Keynesian cross diagram, which is interpreted as follows. Assume a simplified economy where all spending is done by either consumers, investors (the latter, recall, buy capital goods, not stocks and bonds), or governments. (There is assumed to be no trade with other nations, so that net exports are zero, a common simplifying assumption known as the *closed economy* assumption. We briefly discuss the effects of net exports on planned expenditures later in this chapter.) Consumers spend according to the Keynesian (linear) consumption function

$$C = a + bY_D,$$

where C is expenditures by consumers. C depends predominately on income after taxes ($Y_D$) (or disposable income); "b," a parameter known as the *marginal propensity to consume,* gives the rate at which consumption increases in response to a change in disposable income ($0 < b < 1$); "a," a second parameter known as the *autonomous level of consumption,* captures the impact of all forces other than disposable income (such as expectations of the future, household real wealth, etc.) that affect consumption behavior ($a > 0$). It is traditional to convert the consumption function into a function of income before taxes (Y), and taxes, by introducing a simple tax equation into the

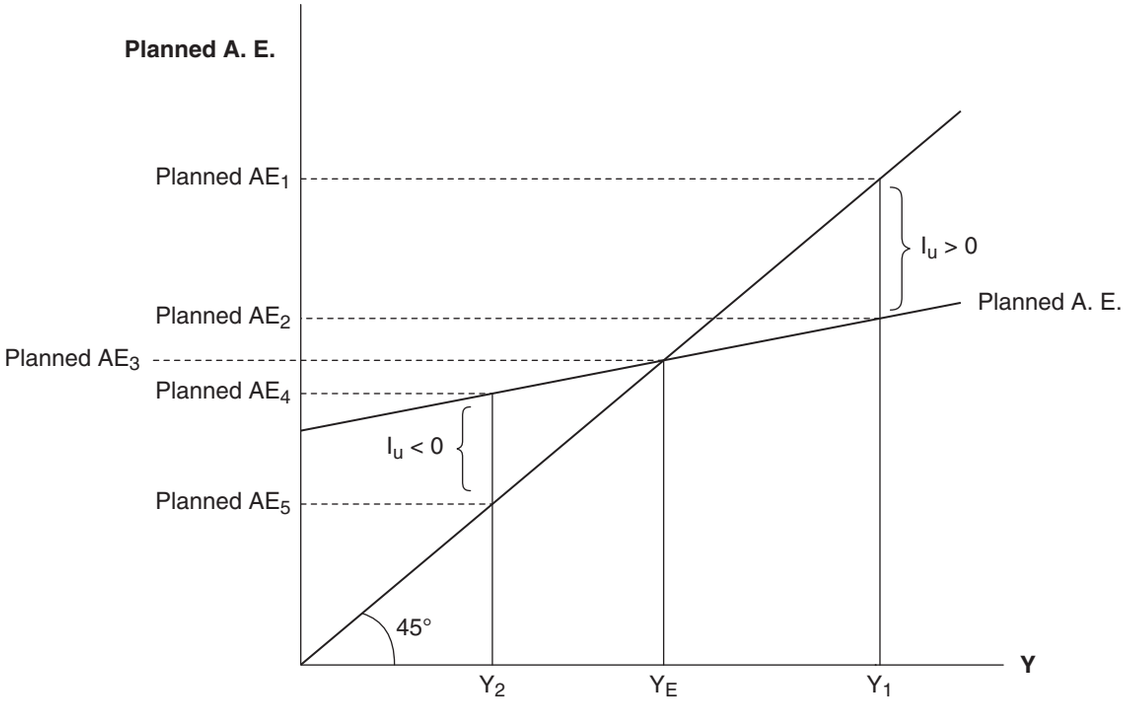


**Figure 31.1** The Keynesian Cross Diagram

discussion. Here we assume a proportional income tax system, so that T equals tY, where T is tax revenues, t is the income tax rate ($0 < t < 1$) and Y is income before taxes.

It is now traditional to assume a linear (or straight-line) consumption function, meaning that, whether income is high or low, the rate at which consumption increases with income remains the same. Keynes himself (1936/1965, pp. 31–32) and some of his followers emphasize the claim that the consumption function is nonlinear, flattening as income rises, so that consumption rises less and less per dollar of increased income as incomes rise. This notion is consistent with the claim that a society with higher income is inclined to spend proportionately less per dollar of new income, making the economy more dependent on investment and government spending if it is to reach full employment. However, many economists do not regard the nonlinearity issue as empirically significant. They prefer to use the linear consumption function, defending it as a reasonable approximation of reality that has the additional advantage of exploiting the flexibility of the linear functional form (vs. the more mathematically complex nonlinear functional form). The use of the linear form has important advantages. For example, strictly speaking, the marginal propensity to consume is defined only for very small changes in disposable income. However, if the consumption function is linear (as is assumed here), the same proportionate relationship holds for small or large changes in disposable income. We assume a linear consumption function throughout this chapter.

Substituting the tax equation into the consumption function and manipulating terms yields the revised consumption function,

$$C = a + b(1 - t)Y.$$

The advantage of this expression for present purposes is that it is expressed in terms that are easily portrayed on the Keynesian cross diagram (which measures real dollars of before-tax expenditure on its vertical axis, and dollars of real income, or output, on its horizontal axis).

The Keynesian cross diagram is designed to highlight, not the determinants of investment spending, but rather the consequences of changes in investment spending and government spending.

Accordingly, to describe the investment sector, let us take the simplest possible description of planned investment spending—one that assumes a given, exogenous level of such spending. Thus, the investment equation is

$$I = I_0,$$

where $I_0$ denotes a given level of planned investment expenditures. The interpretation here of investment spending as planned, not actual, spending is crucial. As described previously, planned investment expenditures are the level of investment expenditures such that businesses expect at the month's close to end up with their planned (or equilibrium) levels of inventory stocks. The interpretation of $I_0$ in this way means that our expenditures expression is

one of planned AE, not actual AE, which, as discussed previously, equals national output by definition. Here it is also useful to recall that planned consumption expenditures equal actual consumption expenditures by definition, and planned government expenditures equal actual government expenditures by definition (the same is true for net exports in versions of the model where they are included).

Our third source of expenditures is government. Again, we make the simplest possible assumption about government spending, which is assumed to be exogenously given as

$$G = G_0.$$

The equation for planned AE in this model is, accordingly,

$$\text{planned AE} = a + b(1 - t)Y + I_0 + G_0 = [a + I_0 + G_0] + b(1 - t)Y.$$

The rightmost side of this expression is what is called in basic algebra the *point-slope form.* Therefore, the intercept term is $[a + I_0 + G_0]$, while the slope of the line equals $b(1 - t)$. When we graph the planned AE expression on the Keynesian cross diagram, the intercept of the planned AE line is where that line intersects the vertical axis. At this intercept point, the value of planned AE equals $[a + I_0 + G_0]$, which is the amount of planned AE that corresponds to a Y level of zero (since people are earning no income at such a point, presumably they are living off their previous savings). Further, the slope of the planned AE line is $b(1 - t)$. Planned AE increases with income due to the consumption function. For every dollar of increase in income, planned AE will increase by $b(1 - t)$. For example, suppose that b equals 0.8 and t equals 0.5. Then $b(1 - t)$ equals 0.4. Each dollar's increase in income will create an additional 40 cents of planned AE (the reverse is true for one dollar's decrease in income). The planned AE line always has a slope less than 1, due to the impact of the marginal propensity to consume and the income tax rate (because both b and t are fractions, their product will also be a fraction). So long as the vertical intercept of the planned AE line is positive (as seems reasonable), the planned AE line will intersect the 45-degree line at some point.

What notion of equilibrium is appropriate for this representation of the macroeconomy? One simple idea is that equilibrium be described as a state where total planned expenditures are just sufficient to purchase all of the economy's output. This means that, in equilibrium, total planned expenditures equal GDP. It also means that planned investment expenditures equal actual investment expenditures. That is, inventories held by firms at the end of the month equal the inventories the firms planned to have when they placed their orders at the start of the month.

Strictly speaking, we are assuming that every firm has zero unplanned inventories. (In applying the framework to the actual economy, the corresponding assumption is that firms on average have inventory levels that are consistent with their planned levels.) All of the goods produced that month are being voluntarily consumed or held by some spender in the economy.

How will this reasoning play out graphically on the Keynesian cross diagram? (Please refer to Figure 31.1.) Notice first the interpretation of the 45-degree line that shoots out from the origin, cutting the graph in two. Because the vertical and horizontal axes are measured in the same units (real dollars of national output), the slope of this line equals 1, so that any point along it is a point where planned spending and national output are equal. That is, any point that rests on the 45-degree line is a point of potential equilibrium in the model. Which particular point is *the* point of equilibrium, however, will be determined by the level of planned AE that is generated by the economy's consumers, investors, and local, state, and national governments. Equilibrium occurs precisely where planned AE intersects the 45-degree line, at the point where planned AE equals planned $AE_3$ and Y equals $Y_E$. Only at this point is the amount of planned spending generated by the economy just equal to the amount of output produced by the economy.

To see further why this intersection is interpreted as a point of equilibrium, consider a point on the planned AE line that is associated with a Y level that is greater than the equilibrium point ($Y = Y_1 > Y_E$). At output level $Y_1$, the amount of planned AE needed to buy up all of the economy's output is given by planned $AE_1$—the planned AE level given by the intersection of the (vertical) $Y_1$ line and the 45-degree line. However, the planned AE actually generated by the economy at $Y_1$ (planned $AE_2$) is given by the intersection of the $Y_1$ line and the planned AE line. The amount of planned AE generated at $Y_1$ is thus less than what is needed for equilibrium.

The deficiency of planned AE at $Y_1$ will reveal itself in the form of unplanned inventory accumulation by firms ($I_U > 0$). ($I_U$ can refer to either unplanned inventory accumulation or unplanned investment spending. This is because all unwanted investment takes the form of unwanted inventories.) Businesses, finding themselves holding more inventories than planned, will reduce their orders of goods. This in turn will reduce production at the factories, pushing the economy's output leftward, away from $Y_1$ toward the equilibrium value $Y_E$. So long as unplanned inventories are positive, this trend will continue until the economy's output reaches the equilibrium point $Y_E$.

Now consider a point on the planned AE line that is associated with output level $Y_2$, which is less than the equilibrium output $Y_E$ ($Y = Y_2 < Y_E$). At output level $Y_2$, the amount of planned spending needed to buy up all of the economy's output (planned $AE_5$) is given by the intersection of the (vertical) $Y_2$ line and the 45-degree line. However, planned AE actually generated by the economy at $Y_2$ (planned $AE_4$) is given by the intersection of the $Y_2$ line and the planned AE line. Planned AE at $Y_2$ is greater

than what is needed for equilibrium. The excess of planned AE at $Y_2$ will reveal itself in the form of unplanned inventory rundown ($I_U < 0$; i.e., negative levels of unplanned inventories) on the part of firms. Firms, finding themselves holding fewer inventories than they had planned, increase their orders of goods. This in turn increases production at the factories, pushing the economy rightward toward higher output levels, away from $Y_2$ toward the equilibrium value $Y_E$. So long as unplanned inventories are negative, this trend will continue until the economy's output reaches the equilibrium point $Y_E$. Actions by firms to move their inventories to equilibrium levels is the primary mechanism driving the economy to its equilibrium on the Keynesian cross diagram.

While the preceding is the standard explanation of how equilibrium is reached on the Keynesian cross diagram, it is not necessary to rely on inventory adjustment to explain the move to equilibrium. In the mid-1980s, Robert Hall and John Taylor (1986) came out with a version of the Keynesian cross model that did not involve inventory accumulation and depletion. Their model, in brief, imagines a world without inventories of goods. Instead, firms "inventory" productive capacity by hiring workers who perform services for firms. An overly optimistic company will find itself with too many workers on staff. Productive capacity then exceeds the actual spending that the economy's spenders are able to generate. The incomes of spenders are insufficient to support the current level of output. This puts downward pressure on the economy and tends to move output downward until the excess productive capacity is eliminated and equilibrium (which Hall and Taylor call *spending balance*) is achieved. Clearly, the model is not too different from the text's inventories-based version, but, as the authors point out, it is useful to show that the mechanisms of the model do not depend on the existence of inventory accumulation in an essential way.

## Applications, Refinements, and Policy Implications

### Comparative Statics of the Keynesian Cross Diagram

The Keynesian cross diagram also can be used to show the impact on AE and equilibrium income of various changes on the demand side of the economy. Starting from the equilibrium position on Figure 31.1, consider, for example, a large, unexpected decline in planned investment. On the Keynesian cross diagram, such a change will shift the planned AE line downward (parallel to the original line). Once planned AE falls, $Y_E$ will no longer be an equilibrium value. Unplanned inventory will accumulate on firms' shelves. They will respond by cutting their factory orders, reducing production at the factories, and eventually bringing about a new equilibrium value at a smaller value of Y (to the left of the former equilibrium value, $Y_E$).

Can anything be done by government to counter such a decline in equilibrium income? In the Keynesian cross framework, the answer is an unambiguous "Yes." The government has the two main tools of fiscal policy at its disposal: changes in government spending (G), and changes in income tax rates (t). By increasing G, the government can shift the planned AE line upward—in principle, by just enough to fully compensate for a decline in planned investment spending. The government may also cut income tax rates, increasing the slope (but not the intercept, given the assumed strictly proportional income tax system) of the planned AE line. The upward rotation by the planned AE line moves the economy's equilibrium Y value upward and rightward along the 45-degree line. Either increases in G or decreases in t, or both, will increase equilibrium income, through our now standard mechanism of inducing changes in inventory accumulation (either causes inventory rundown). Retailers then increase their orders, increasing production in the factories and raising Y. The Keynesian cross diagram neatly illustrates the potential for activist government fiscal policies to counter undesirable changes in private sector spending.

Monetary policy's impact also may be represented on the Keynesian cross diagram. A stimulative monetary policy—one that increases the money supply or its rate of growth—is commonly viewed as lowering the economy's interest rates, which in turn triggers additional borrowing by firms, who then spend the money on new investment projects. The resulting rise in planned investment expenditures can be interpreted on the Keynesian cross diagram as an upward shift in the planned AE curve, which then stimulates the economy through the same inventory-based adjustment mechanism.

### The "Open" Economy

The preceding discussion, for simplicity, assumes a *closed economy*—that is, an economy that does not trade with other nations. How do things change when trade with the rest of the world is introduced? Now, a fourth potential source of spending (which can be either positive or negative) is introduced: net exports. Net exports equal the goods and services that the economy exports to other nations of its production (exports), minus the goods and services that the economy imports from abroad (imports). When exports are greater than imports, then net exports are positive and planned AE is larger than in the case of the closed economy (the economy, in this case, is often said to be running a *trade surplus*). Citizens of other nations are adding to our economy's domestic demand, causing planned AE to rise. Accordingly, when net exports are positive, the planned AE line shifts upward on the Keynesian cross diagram. By contrast, when exports are less than imports, then net exports are negative and planned AE is smaller than in the closed-economy case (the economy, in this case, is often said to be running a *trade deficit*). In this case, our economy, on net, is buying goods from abroad

that often could be purchased domestically. Other things held constant, this means less demand for goods produced in our economy. Accordingly, when net exports are negative, the planned AE line shifts downward on the Keynesian cross diagram.

When net exports are included as a source of AE, the model is known as an *open economy model.* In the open economy, AE depends on the volume of net exports in two fundamental ways. First, an increase of income in the "home" country tends to increase the purchases of goods generally, including goods produced abroad. In the short run, this increase in home country spending increases imports into the home country without increasing exports (as incomes abroad are as yet unchanged). Thus, net exports decrease and, other things equal, planned AE in the home country declines. (In the event of a given fall in income, the effect is reversed.)

Second, a given increase in interest rates in the home country relative to abroad encourages a flow of investable funds into the home country. The process of investing these funds in the home country drives up the home country's exchange rate, making home country goods more expensive to foreign citizens, and making foreign goods cheaper for home country citizens. Thus, exports fall, imports rise, and net exports fall, reducing the level of AE in the home country (other things equal). (In the event of a given fall in interest rates, the effect is reversed.)

## Meaning of "Equilibrium" in the Keynesian Cross Diagram

The interpretation of the so-called equilibrium on the Keynesian cross diagram has been the subject of much controversy. Much of the controversy centers on the concept of *full employment* and the question of whether or not an economy could be in equilibrium without being at full employment. When economists speak of the full-employment position of the economy, they mean the economy's maximum production level that is consistent with macroeconomic well-being. Too little production (i.e., an economy below full employment) means that the economy is not getting all the output out of its scarce resources that it should. Too much production (i.e., an economy above full employment) means that the economy is overheated. That is, the economy is producing at such a hefty pace that painful side effects like inflation and spot shortages of crucial raw materials (known as *bottlenecks*) are being created to an unhealthy degree. Above full employment, it is also possible for an economy to find itself destabilized, triggering a sharp decline in economic activity (even an outright recession).

In Chapter 3 of the *General Theory,* Keynes presented an aggregate demand and aggregate supply framework that later inspired the Keynesian cross diagram. Keynes emphasized that there was no particular reason why the equilibrium reached by the economy in his framework should correspond with full employment. Equilibrium could be above, equal to, or below full employment (though Keynes was quite sure that it would usually be below it; 1936/1965, pp. 118, 254). Keynes saw this equilibrium as a true point of stability—once reached, it would be hard for the economy to move itself away from it and to the economy's full employment position. To classically oriented macroeconomists who criticized Keynes, the notion of a stable macroeconomic "equilibrium" that was not at full employment was almost a contradiction in terms. (A lengthy debate about this and related matters dominated macroeconomics through much of the late 1930s and 1940s.)

When followers of Keynes such as Alvin Hansen and Paul Samuelson substituted the Keynesian cross diagram for Keynes's framework, they kept Keynes's sharp separation of equilibrium and full employment. Over time, however, macroeconomists grew increasingly uncomfortable with a notion of macroeconomic "equilibrium" that was not also situated at full employment (see, in particular, Patinkin, 1948, esp. pp. 562–564). As it had been before Keynes, price flexibility—a slow but still steady and substantial force in the economy—became again a generally accepted mechanism through which economies that were away from full employment were returned there. Many of Keynes's successors downgraded the Keynes/Hansen/Samuelson notion of equilibrium to the lower status of a mere *short-run equilibrium. Long-run equilibrium* was then defined in the traditional, classical economics manner as something occurring only at full employment.

The idea behind this divergence of equilibrium notions probably came from the microeconomics of Alfred Marshall. Marshall taught Keynes and doubtless influenced him in his decision to propose a short-run equilibrium concept in macroeconomics. In Marshall's microeconomic theory of perfect (or pure) competition, the short-run equilibrium is defined for a case where the number of firms in the industry is fixed (there is not enough time for firms to enter or exit the industry). In the long run, when free entry and exit are allowed, a long-run equilibrium emerges that is quite distinct from the short-run equilibrium.

The macroeconomic version of the idea is that, while there are certain slow-acting forces that eventually move the economy to full employment, they are dominated in the short run by quicker acting forces. Chief among these slower acting forces is price level adjustment. In theory, it alone is sufficient to eventually move an economy to the full employment position. However, in the short run, the economy's price level moves little, so that it can safely be assumed to be constant (or "sticky"). Based on such reasoning, a macroeconomic short-run equilibrium exists that is reached not through price changes but rather through changes in quantities—such as the inventory adjustment mechanism described previously. Thus, the argument concludes, a short-run model ought to emphasize planned AE and its determinants and place little emphasis on full employment, while a long-run model of the economy ought to emphasize full employment and to de-emphasize traditional Keynesian notions in favor of more "classical" ones.

More recently, the equilibrium notion in the Keynesian cross diagram has been downgraded even further into a mere "spending balance" (see Hall & Taylor, 1986, or a more recent edition of the text, such as Hall & Papell, 2005). Hall and Taylor divorce the model from its long-standing equilibrium connotation, even to the point of denying to it the term itself. The idea is that the intersection of planned AE and the 45-degree line represents a balancing point for the economy, which the economy reaches quite quickly, but which should not really be conceived as an equilibrium per se because the balance point changes quickly whenever the level of (planned) expenditures alters. Hall and Taylor reserve the term *equilibrium* only for the full employment position, insisting that only after the price adjustment process has moved the economy to the full employment point can the notion of equilibrium be usefully invoked. The authors' approach to interpreting the intersection of the 45-degree line and the planned AE line is in the spirit of modern treatments, which insist that the healing role played by price adjustment in macroeconomics be taken seriously, not suppressed, as in the Keynesian cross model.

Here it should be pointed out that Keynes and his immediate followers were thoroughly skeptical of aggregate price adjustment as a cure for recession. Keynes felt strongly enough about the issue to present a detailed critique of it as his first major topic in the *General Theory.* In a nutshell, Keynes and his followers emphasize how lower input prices brought about by price adjustment adversely impact incomes of those already employed. This downward push to demand could more than wipe out the supply-side gains from lower costs of production. For details, see Keynes (1936/1965, chaps. 2, 3, 20). Modern macroeconomics, as portrayed in the *Principles* textbooks (e.g., McConnell, 1969), implicitly assumes that these adverse demand-side forces are dominated strongly by supply-side effects of lower input prices. Later versions of Keynesianism present arguments backing up the assumption (see, e.g., Snowdon & Vane, 2005, pp. 114–116).

## Strengths and Weaknesses of the Keynesian Cross Model

Since its origination in the 1930s and 1940s by Keynes, Hansen, and Samuelson, the Keynesian cross model has been a durable workhorse of macroeconomics. Its primary strength is its highlighting of inventory adjustment as a crucial part of the short-run aggregate adjustment process. There is little doubt that incomplete inventory adjustment and the changes in output that such incomplete adjustment calls forth makes a significant contribution to short-run aggregate fluctuations. The financial press routinely refers to an inventories-driven business cycle that has contributed to cyclical movements in the post–World War II era (see, e.g., Gramm, 2009). The Keynesian cross framework captures these important real-world forces neatly in one tight diagram. As such, it is a useful tool for teaching about one of the core short-run adjustment processes of the macroeconomy (inventory adjustment) and its aggregate consequences.

That having been said, the model is not without its weaknesses. The most glaring is that prices in the model—including inventory prices—are fixed exogenously outside of the model. Consequently, one of the most fundamental mechanisms in economics—the tendency of surpluses to be eliminated by price declines, and vice versa—is not allowed to function in the Keynesian cross framework. Moreover, there is no link between the inventory adjustment process and aggregate price determination in the model. The constant-price assumption has been a significant, and much criticized, feature of many Keynesian models (particularly those dating from the early years of Keynesianism). The Keynesian cross diagram is an excellent example of such a model. Keynesians can, and do, respond that the time period of analysis assumed in the Keynesian cross model is too short to allow meaningful price adjustment. Keynesians can also cite plausible reasons why various prices should be "sticky" over short periods of time. However, the notion that price adjustment does not play at least some role in firms' elimination of stocks of unplanned inventories is, to many economists, counterintuitive. The best that can be said on this point is that the Keynesian cross model should be conceived as an exploration of the aggregate consequences of inventory adjustment subject to the assumption that prices are fixed. (Some additional criticisms of the Keynesian cross framework are discussed in Weintraub, 1977, and Guthrie, 1997.)

This criticism applies mainly to the Keynesian cross diagram as a stand-alone model. The model works better when conceived as an "input" into what Hall and Taylor (1986) have called the *aggregate demand/price adjustment model.* In this framework, the role of the Keynesian cross diagram is mainly to help generate the so-called IS curve, which gives an equilibrium relationship between various interest rate levels and planned AE. The IS (or "investment = savings") curve gives all the equilibrium points relating the interest rate and real income for a given set of model "parameters." It is drawn on a graph with the interest rate on the vertical axis and the level of real income (or real output) on the horizontal axis. The IS curve is downward sloping, reflecting how a higher interest rate drives down planned investment spending and thus lowers the economy's equilibrium income (and vice versa). The IS curve is commonly derived graphically from the Keynesian cross diagram. On the Keynesian cross diagram, a rise in the interest rate shifts the planned AE line downward (and vice versa). The derivation is done as follows. First, make an empty graph with the interest rate on the vertical axis and real income (or real output) on the horizontal axis. Second, place that graph directly above a Keynesian cross diagram. Because real output is on both horizontal axes, the two graphs "line up" vertically. Third, vary the interest rate, observing the

resulting values in equilibrium income that are generated on the Keynesian cross diagram. Finally, derive the IS curve by plotting this relationship between the interest rate and equilibrium income onto the upper graph.

Once it is derived, the IS curve is combined with a Liquidity = Money Supply (LM) curve and a price adjustment equation to create a general model of the expenditures side of the macroeconomy. The IS/LM graph in turn can be used to generate an aggregate demand curve, which relates various values of the price level to planned AE. Then, this aggregate demand curve is combined with a supply-side model to give a general macroeconomic model of the aggregate economy. This model incorporates everything from the macroeconomic consequences of deviations in planned investment spending (including planned inventory spending) and other demand-side changes, to the macroeconomic consequences (and causes) of supply-side forces and price level changes. Numerous additional important topics can be analyzed as well within this broader framework.

## Historical Development of the AE/Keynesian Cross Framework

### Development of the Planned AE (Aggregate Demand) Concept

While the Keynesian cross framework emerged out of the Keynesian revolution, many of the ideas behind that model are far older than Keynesianism. Planned AE was called *effectual demand, effective demand,* or *purchasing power* in pre-Keynesian days. Even as early as the eighteenth century, the mercantilist school emphasized what Keynes (1936/1965) later called "the great puzzle of effective demand" (p. 32; see also his chap. 23). Most leading classical economists downplayed the notion of deficiencies in planned expenditures (or aggregate demand, as planned expenditures were usually called at the time). They believed that, except for brief and self-correcting "crises," planned AE would take care of itself so long as production was unfettered and the types of goods produced were consistent with what consumers wanted and were able to purchase. Thomas Malthus, a notable exception, insisted that an economy could have serious drawn-out problems with inadequate effective demand (his arguments were a regular bone of contention in the early nineteenth century between himself and David Ricardo, a key founder of classical economics). Keynes lauds Malthus's position and writes also of how, during the heyday of the classical school, ideas that took effective demand failure seriously "could only live on furtively, below the surface, in the underworlds of Karl Marx, Silvio Gesell or Major Douglas" (p. 32).

Like Keynes, early anticlassical thinkers were concerned mainly with the prospect of secular (long-term) demand failure—not just with occasional downward fluctuations in planned AE that contributed to a downward swing in the business cycle. Keynes saw the laissez-faire macroeconomy as routinely—not just occasionally—underperforming its potential due to problems with generating sufficient levels of planned expenditures (see, e.g., Keynes, 1936/1965, chap. 1; chap. 3, p. 33, bottom paragraph; chap. 12, p. 164, final sentence; chap. 18, p. 254). There is also the fact that Keynes chooses to develop a business cycle theory only at the end of the book (chap. 22). If the *General Theory* were primarily about combating the business cycle, purely cyclical forces surely would have been given a more prominent role in the main portion of the volume.

Keynes's (1936/1965) chief (though far from sole) culprit causing inadequate AE was unstable confidence by the business investment sector, which created secular underinvestment and thus routine unemployment problems (chap. 12). The worldwide Great Depression, which raged even as Keynes was writing his famous *General Theory,* seemed thoroughly consistent with Keynes's analysis. Never before had the hypothesis of secular effective demand failure seemed so confirmed by day-to-day events.

This certainty, however, faded with time in the post–World War II era in the face of a recovered economy and a withering fire by the inheritors of the classical tradition. Since Keynes wrote, the "market failure" interpretation of the Great Depression has been effectively challenged by the heirs of the classical tradition. The most famous example is Milton Friedman and Anna Schwartz (1963; see also Friedman & Friedman, 1980, chap. 3), which emphasizes an unprecedented contractionary monetary policy during the crucial 1929 to 1933 period. Others have focused on the extraordinary uncertainty created by government actions during the decade of the 1930s (see Anderson, 1949; Smiley, 2002; Vedder & Gallaway, 1993).

It is a common misconception that the classical economists of the late eighteenth and early nineteenth centuries, as well as the early twentieth-century heirs of the classical tradition, had no interest in, or theory of, planned AE. Indeed, this is arguably a fair assessment of some classical writers. However (though the term itself was not used), short-term declines in planned aggregate spending were seen by several classical economists as being a key part of every business cycle downturn (or *crises,* as they were called at the time). Classical economists, however, differed from Keynes in two significant ways. First, they flatly rejected any notion of secular failure in planned AE. Second, they did not see declines in planned expenditures as being the ultimate cause of recession. Instead, they saw these declines as being caused by other more fundamental factors (such as disruptive government regulatory, or monetary, policies). Even before the classical era, David Hume (1752/1955, pp. 37–40) presented a very modern tale of a recession being caused by an unexpected decline in the money supply. Nearly a century later, John Stuart Mill (1844/1948, esp. pp. 68–74) described how a crisis of confidence could cause people to hoard money and so reduce their purchases of goods. Allowing for the difference in

vocabularies between his time and ours, Mill accurately described such an episode as being caused by a short-term decline in planned AE (or aggregate demand).

Early twentieth-century macroeconomists (whom Keynes, peculiarly, also labeled *classical*) also focused on aggregate demand issues—though always in the context of trying to understand cyclical, rather than secular, fluctuations. Leland Yeager (1997) summarizes numerous early twentieth-century business cycle theories that accurately identified the demand-side forces at play in the downturn—although such a downturn in demand was seen routinely as the effect of still more basic forces (such as a decline in the money supply). J. Ronnie Davis (1971) details how economists in the 1930s at the University of Chicago and elsewhere pursued lines of thinking parallel to Keynes's, suggesting that Keynes's ideas were "in the air" at the time.

Still, despite these glimmers from earlier scholars, it is without a doubt John Maynard Keynes who is primarily responsible for the ascendancy of demand-side macroeconomics through the influence of his *General Theory* (1936/1965). Keynes confronted a profession that took little explicit interest in the study of aggregate demand and left that profession positively consumed by the topic. Short-run macroeconomics, including the study of the business cycle, has since Keynes's day placed great emphasis on questions pertaining to short-run planned AE. Through about 1976, the primary focus of macroeconomics was to examine the theory and determinants of planned AE (Friedman, 1968, is still the classic brief summary of much of the period covered in this section. It is highly recommended to the student's attention). The main determinants of planned AE—consumption, investment, government, net exports, and the demand for and supply of money—the statistical properties of these components, and ways for the monetary and fiscal authorities to alter these aggregates are topics that have all been extensively explored. The profession's comprehension of the practicality of the basic Keynesian policy agenda has improved markedly as all aspects of that agenda have been placed under the magnifying glass.

At the risk of some oversimplification, it may be said that two fundamental issues dominated the policy agenda of those who worked on developing the Keynesian framework in the 30 years following Keynes's death in 1946. The first concerned mainly the channels and mechanisms through which demand-side policy worked, while the second focused on the relative strengths and weaknesses of fiscal policy (changes in government spending, tax-rate changes) versus monetary policy (changes in the money supply or in its rate of growth). Through this entire era, understanding was enhanced considerably by the development of statistical methods in economic science. Nearly every significant concept developed by Keynes that pertained to the control of planned AE was exhaustively examined, from both theoretical and statistical perspectives. (Snowdon & Vane, 2005, chaps. 1–7, is among the most useful detailed reviews of the material summarized here.)

For example, how exactly did policy and consumption interact in generating economic stimulus? Keynes (1936/1965) confined his analysis of consumption mainly to its relationship with current disposable income (chaps. 8–10). In fact, as Milton Friedman (1957) and others soon showed, expected future wealth plays a crucial role as well, especially in policy matters. Friedman showed that one recession-fighting policy tool—income tax cuts aimed at consumers—would tend to stimulate significantly less spending than that implied by the Keynesian consumption function. Consumers, looking to the future, would observe the temporary nature of the tax cut (rates would be raised again as the economy recovered). They would therefore seek to spread out the spending of their income windfall over many future periods rather than spend the bulk of it in the period of the tax cut. The result was to limit the relevance of income tax cuts as a demand-stimulating mechanism that might fight recessions. Income tax systems seem to work better as automatic stabilizers, because as income falls, one's tax burden automatically is lowered also (and vice versa) in such tax systems. But as a means of direct assault on recession, the case for such cuts, when viewed as demand-side stimulus, appears weaker than was believed in the early Keynesian era, thanks to the work of Friedman and others (specifically, Duesenberry, 1948; Modigliani & Brumberg, 1954). (So-called supply-side economics advocates income tax cuts as a means of stimulating work effort, reducing income tax avoidance/evasion, and stimulating entrepreneurial small-business activity. Friedman's work did not refute these claims. However, these are not Keynesian arguments. The supply-side argument for income tax cuts remains controversial.)

A number of additional insights were developed in the 30 years following Keynes's death. The inverse relationship between investment spending and interest rates, which had been controversial, was established as fairly strong and reliable as better statistical methods became available. Investment spending's interaction with income (known as the *accelerator mechanism*) also was established. In addition to Keynes's basic mechanism, several significant channels through which monetary policy affected planned aggregate spending were identified, and the potential for changes in the demand for money to complicate monetary policy was thoroughly examined. The determinants of net exports became much better understood. Finally, complicating factors affecting the potency of both fiscal and monetary policy became far better understood. In particular, the existence of lags delaying the effectiveness of such policies came to be seen as a significant complicating factor, especially in the case of major stimulus initiatives that involved substantial increases in government spending in an attempt to fight recessions. Invariably, it seems, increases in government expenditures to fight recessions add little stimulus until well after the recession has passed (this has, at least, been the lesson of the post–World War II era up to the present date). Lags also afflict monetary policy, although it appears that monetary policy is more able

to retain a significant stimulative role within a time frame capable of combating cyclical forces.

As understanding of the determinants of planned AE increased, a sharp debate developed between advocates of traditional Keynesian fiscal policy measures and those who advocated the superiority of monetary policy over fiscal policy (these latter scholars, who were led by Milton Friedman, were known as *monetarists,* and their movement as *monetarism*). Keynes, in the *General Theory,* had focused attention on expansionary fiscal policies as the primary cure for demand-side problems. Keynes had not ignored monetary policy, and he endorsed it also, but he feared it would be weak precisely when it was needed most—in depressed economic times. Some of Keynes's early followers went so far as to nearly deny to monetary policy any significant impact on planned AE and real GDP. This pessimism proved overblown; in fact, one of the major themes of the 1960s and 1970s was the steady growth of a consensus among macroeconomists that monetary policy was the more potent of the two policy "levers."

Another insight painfully developed during the initial post-Keynes era was a clear-cut link between monetary policy and inflation. By the mid-1970s, inflationary forces (largely stimulated by loose monetary policy during the late 1960s and early 1970s) dominated the policy front. A common notion that there was an exploitable policy trade-off between higher inflation and lower unemployment (a trade off that became known as the *Phillips curve*) was shattered by the high inflation and high unemployment of the late 1970s and early 1980s. Out of this era, for a time, there emerged a deep cynicism about macroeconomic policy generally—both fiscal and monetary, though the focus at this time inevitably was on monetary policy, which was widely seen as the more potent of the government's two core policy choices. In particular, advocates of the so-called rational expectations hypothesis argued that monetary policy would have little impact on planned AE unless that policy was unanticipated. Otherwise, "rational" individuals would alter their personal economic affairs in such ways as to defeat the stimulative intent of the expansionary monetary policy.

By the mid-1980s, the rational expectations critique of monetary policy had been exposed as overly extreme (even as the rational expectations hypothesis itself became a standard feature of macroeconomic theorizing) due to the clear-cut potency of monetary policy during the 1980s (and beyond) to impact planned AE and real GDP. During this era, attention turned to supply-side theories of cyclical activity (so-called real business cycle theory). Real business cycle theorists aggressively challenged traditional Keynesian demand-based theory, claiming (in essence) that such theories were internally inconsistent. *New Keynesian economics* developed in response to this fundamental challenge. By the middle of the 1990s, new Keynesians had met the challenge posed by real business cycle theory, in the process developing several new mechanisms through which monetary and fiscal policies might impact planned AE and real GDP. The new Keynesian success at shoring up the theoretical foundations of their macroeconomic approach had the effect of reestablishing the influence of traditional Keynesian macroeconomic analysis, which had suffered in reputation during the 1975 to 1985 period.

At least in mainstream macroeconomics, there has been no significant new challenge to the supremacy of AE policies since the real business cycle movement in the 1980s (outside the mainstream, "Austrian" macroeconomics questions this supremacy; see, e.g., Snowdon & Vane, 2005, chap. 9). Indeed, during the 1990s and early 2000s, the potency of monetary policy to stabilize an unstable economy never looked brighter. Several severe financial crises in the late 1990s and early 2000s seemed largely blunted by expansionary monetary policy, aided to an extent by expansionary fiscal policies (mainly tax cuts). However, as the first decade of the twenty-first century nears its end, unprecedentedly severe disruptions to the world's financial sector have called again to the fore the core question of the potency of traditional stabilization policy measures to combat a truly severe economic contraction. It is to be hoped that the many insights achieved since Keynes's death will prove sufficient to meet the phenomenon of the burst of a worldwide speculative bubble—a specter deeply feared and much discussed by pre-Keynesian economists, but an issue relatively unexamined in macroeconomics since 1946.

## Development of the Keynesian Cross Diagram

As is the case for planned AE, aspects of the Keynesian cross diagram also can, in part, be attributed to classical thinkers. John Stuart Mill (1844/1948) in particular displays a thorough comprehension of the inventory adjustment mechanism at the heart of the Keynesian cross. Mill writes about how in bad economic times "dealers in almost all commodities have their warehouses full of unsold goods," and how under these circumstances "when there is a general anxiety to sell and a general disinclination to buy, commodities of all kinds remain for a long time unsold" (pp. 68, 70). Seventy years later, Ralph Hawtrey (1913/1962, p. 62) shows a particularly clear grasp of the relation between inventory accumulation/depletion and aggregate economic activity. Classical economists and heirs to the classical tradition may not have conceptualized aggregate demand in the manner of Keynesian economics, but they understood the concept and its significance well enough, particularly in relation to the business cycle.

The history of the Keynesian cross framework proper begins, of course, with Keynes. Like Mill before him, Keynes (1936/1965) saw clearly the role of inventory accumulation and depletion in contributing to cyclical forces. In his Chapter 22 in the *General Theory* on the "Trade Cycle," Keynes points to how, in the downswing, the "sudden cessation of new investment after the crisis will probably lead to an accumulation of surplus stocks of

unfinished goods." He then comments, "Now the process of absorbing the stocks represents negative investment, which is a further deterrent to employment; and, when it is over, a manifest relief will be experienced" (p. 318). This is precisely the message of the Keynesian cross diagram.

However, Keynes introduces the framework that would give birth to the Keynesian cross diagram much earlier in the *General Theory.* In Chapter 3, he verbally describes a model where effective demand is set by the intersection of aggregate supply and aggregate demand. The specifics of the model need not detain us, but the model featured an upward-sloping aggregate supply relationship between employment and the minimum proceeds businesses must receive to hire a given number of workers (aggregate demand is given by what businesses actually receive in proceeds for hiring a given number of workers). Keynes does not draw a graph, but he does assume that the two schedules intersect, and under reasonable conditions his aggregate demand curve will be flatter than the aggregate supply curve.

Samuelson (1939, p. 790) was apparently first to publish a version of Keynes's framework featuring the now-standard 45-degree line combined with a flatter C + I curve representing planned AE. Samuelson also pioneered the teaching of the Keynesian cross framework by including it in early editions of his best-selling *Principles of Economics* textbook. Some early interpretations of the 45-degree line treated it as a kind of aggregate supply line (see Guthrie, 1997, for a useful discussion), and this interpretation also found its way into the *Principles* textbooks (e.g., McConnell, 1969). However, the more standard approach has been the modern one of treating the 45-degree line simply as a line of reference that shows all points where a value on the vertical axis equals that on the horizontal axis (as in Dillard, 1948; Samuelson, 1939).

## Conclusion

The development and refinement of the planned AE (or aggregate demand) concept has been a dominant theme of macroeconomics since at least 1936. Keynes (and, to a surprising extent, earlier "classical" thinkers) forged the initial path, and by 1976, a sophisticated and intricate theory of planned AE had been developed. This development has continued into the twenty-first century under the spur of attacks by opposing schools of thought that came to the fore in the late 1970s and 1980s. While challenges have been raised from time to time, it is still safe to say, 10 years into the twenty-first century, that countercyclical macroeconomic policy is primarily defined by the emphasis on planned AE that was put into place by Keynes nearly a century ago.

Few macroeconomic models have received more attention, or been studied by more college sophomores, than the Keynesian cross diagram. The model's chief virtue always has been the neat—even elegant—manner in which the Keynesian approach to macroeconomics can be made clear. The causes and effects of sudden fluctuations in planned AE can be easily and instructively represented in the model. Demand-side policy solutions recommended by Keynesian economics also are easily highlighted in the model. The model captures well one of the chief dynamics of macroeconomic adjustment—inventory accumulation and depletion, and its consequences. It also dovetails nicely (indeed, it is an input into) the more complete IS/LM model of planned expenditures, which, when combined with a theory of aggregate price adjustment, is a creditably complete, and remarkably informative, aggregate model of the macroeconomy.

If the Keynesian cross framework has one serious weakness, it is that it encourages the student to accept far too easily that AE can be easily and conveniently manipulated by governments in such a way as to improve the aggregate performance of the economy. The model presents macroeconomic problems and their solutions in a commendably clear, but vastly oversimplified, manner. It is all too easy for students to emerge from courses built around the Keynesian cross with an overblown faith in the government's ability to move an economy briskly and with almost surgical precision to "macroeconomic nirvana." In fact, lags in the effectiveness of macroeconomic policies, political complications that "hijack" such policies for the sake of personal interests, expectations-based problems where policies work only when implausible beliefs about them are held by people, and other problems bedevil the macroeconomic policy maker. Still, when the model is taken with a grain of salt and on its own terms, it can do an admirable job of acquainting the student with the primary motivating ideas of Keynesian economic theory.

## References and Further Readings


Anderson, B. M. (1949). *Economics and the public welfare: Financial and economic history of the United States, 1914–1946.* Princeton, NJ: D. Van Nostrand.

Davis, J. R. (1971). *The new economics and the old economists.* Ames: Iowa State University Press.

Dillard, D. (1948). *The economics of John Maynard Keynes: The theory of a monetary economy.* New York: Prentice Hall.

Duesenberry, J. S. (1948). *Income, saving, and the theory of consumer behavior.* Cambridge, MA: Harvard University Press.

Friedman, M. (1957). *A theory of the consumption function.* Princeton, NJ: Princeton University Press.

Friedman, M. (1968). The role of monetary policy. *American Economic Review, 58,* 1–17.

Friedman, M., & Friedman, R. (1980). The anatomy of crisis. In *Free to choose: A personal statement* (pp. 62–81). New York: Harcourt.

Friedman, M., & Schwartz, A. (1963). *A monetary history of the United States, 1867–1960.* Princeton, NJ: Princeton University Press.

Gramm, P. (2009, February 20). Deregulation and the financial panic. *Wall Street Journal,* p. A17.

Guthrie, W. (1997). Keynesian cross. In T. Cate (Ed.), *An encyclopedia of Keynesian economics* (pp. 315–319). Cheltenham, UK: Edward Elgar.

Hall, R. E., & Taylor, J. B. (1986). *Macroeconomics: Theory, performance, and policy.* New York: W. W. Norton.

Hall, R. E., & Papell, D. H. (2005). *Macroeconomics: Economic growth, fluctuations, and policy* (6th ed.). New York: W. W. Norton.

Hansen, A. H. (1953). *A guide to Keynes.* New York: McGraw-Hill.

Hawtrey, R. (1962). *Good and bad trade: An inquiry into the causes of trade fluctuations.* New York: Augustus M. Kelley. (Original work published 1913)

Hume, D. (1955). Of money. In E. Rotwein (Ed.), *Writings on economics* (pp. 33–46). Edinburgh, UK: Nelson & Sons. (Original work published 1752)

Keynes, J. M. (1965). *The general theory of employment, interest and money.* New York: Harbinger. (Original work published 1936)

McConnell, C. R. (1969). *Economics: Principles, problems and policies* (4th ed.). New York: McGraw-Hill.

Mill, J. S. (1948). On the influence of consumption upon production. In J. S. Mill (Ed.), *Essays on some unsettled questions in political economy* (Series of Reprints on Scarce Works in Political Economy No. 7; pp. 47–74). London: London School of Economics and Political Science (University of London). (Original work published 1844)

Modigliani, F., & Brumberg, R. (1954). Utility analysis and the consumption function: An interpretation of cross-section data. In K. K. Kurihara (Ed.), *Post-Keynesian economics* (pp. 388–436). New Brunswick, NJ: Rutgers University Press.

Patinkin, D. (1948). Price flexibility and full employment. *American Economic Review, 38*(4), 543–564.

Samuelson, P. A. (1939). A synthesis of the principle of acceleration and the multiplier. *Journal of Political Economy, 47*(6), 786–797.

Smiley, G. (2002). *Rethinking the Great Depression.* Chicago: Ivan R. Dee.

Snowdon, B., & Vane, H. R. (2005). *Modern macroeconomics: Its origins, development and current State.* Cheltenham, UK: Edward Elgar.

Vedder, R. K., & Gallaway, L. E. (1993). *Out of work: Unemployment and government in twentieth-century America.* New York: Holmes & Meier.

Weintraub, S. (Ed.). (1977). *Modern economic thought.* Philadelphia: University of Pennsylvania Press.

Yeager, L. B. (1997). New Keynesians and old monetarists. In L. Yeager (Ed.), *The fluttering veil: Essays on monetary disequilibrium* (pp. 281–302). Indianapolis, IN: Liberty Fund.

# 32

# Aggregate Demand and Aggregate Supply

Ken McCormick
*University of Northern Iowa*

The aggregate demand/aggregate supply (AD/AS) model appears in most undergraduate macroeconomics textbooks. In principles courses, it is often the primary model used to explain the short-run fluctuations in the macroeconomy known as *business cycles*. At the intermediate level, it is typically linked to an IS/LM model. The IS/LM model is a short-run model, and in intermediate macroeconomics classes, the AD/AS model serves as a bridge between the short run and the long run. Many economists find the model to be useful in thinking about the macroeconomy.

People often assume that textbooks present settled theory. It might therefore be a surprise to some that economists do not universally accept the AD/AS model. It has, in fact, been the subject of a heated debate. Critics of the model charge that it is an unsuccessful attempt to combine Keynesian and neoclassical ideas about the macroeconomy. Detractors also contend that it sacrifices accuracy for the sake of easy pedagogy.

The purpose of this chapter is to examine the model and a few of the charges leveled against it. The first part of the chapter presents the model as it usually appears in textbooks. The presentation includes applications and a discussion of policy in the context of the model. Following that is a discussion of some of the main criticisms. We discover that the critics may have a point.

## Theory

### The Aggregate Demand Curve

The aggregate demand curve is drawn as a negatively sloped curve in price level/real gross domestic product (GDP) space (see Figure 32.1). It shows how real aggregate spending is influenced by changes in the price level. Aggregate spending is the sum of consumption, desired investment, government purchases of goods and services, and net exports, symbolized as C + I + G + NX. At the principles level, there are usually three explanations given for why real aggregate spending varies inversely with the price level: the wealth effect, the interest rate effect, and the foreign trade effect.

The wealth effect is also known as the *real balance* or *Pigou effect*. All else equal, as the price level rises, people

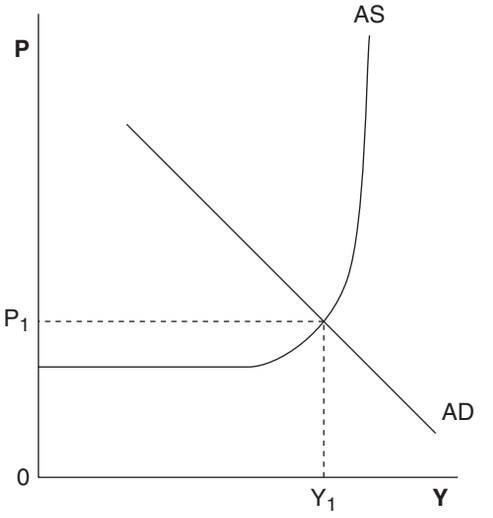


**Figure 32.1** The Basic AD/AS Model
SOURCE: Adapted from Reynolds (1988, p. 221).

who hold money discover that they can buy less with their money. In other words, they experience a decrease in the real value of their wealth held as money. As a direct result, they will buy fewer goods and services as the price level rises.

The interest rate effect is sometimes called the *Keynes effect.* To understand it, one must first understand the theory of *liquidity preference,* which is the idea that in the short run, interest rates are determined by the supply of and demand for money. In the money market, the nominal supply of money is determined by the central bank. The nominal demand for money depends on real income, the price level, and the nominal interest rate. All else equal, as real income rises, people will buy more goods and services. To do so, they will need more money. As the price level rises, people will need more money to carry out the same volume of real transactions. Money demand is negatively related to the interest rate because the rate of interest is the opportunity cost of holding money. All else equal, as the interest rate rises, people will reduce the amount of money they want to hold.

The *equilibrium short-run interest rate* is the rate that causes the quantity of money demanded to equal the quantity of money supplied. If the interest rate is too high, there will be a surplus of money in the sense that people will want to hold less money than they do. To reduce money holdings, people will buy bonds and deposit money in interest-bearing accounts. The result will be a fall in the rate of interest. If the interest rate is too low, the opposite will occur. People will want to hold more money than they have. To obtain it, they will sell bonds and withdraw money from interest-bearing accounts, and that will drive the interest rate up.

The liquidity preference theory is in contrast to how interest rates are determined in the long run. In the long run, interest rates are determined in the market for loanable funds—that is, by national saving, domestic investment, and net capital outflow (also known as *net foreign investment*). According to the doctrine of long-run money neutrality, the supply of money cannot influence the real interest rate in the long run.

So what is the interest rate effect and how does it affect the slope of the aggregate demand curve? There are two different approaches presented in textbooks that are logically equivalent. In one approach, an increase in the price level increases the nominal demand for money. For a given money supply, the increase in money demand drives up the rate of interest. A higher interest rate reduces consumption and desired investment spending. In other words, a higher price level drives up the rate of interest and thereby reduces the real quantity of output demanded.

The other approach conceives of the money market in real terms. By dividing the nominal money supply and nominal money demand by the price level, one gets real money supply and real money demand. Holding the nominal money supply constant, as the price level rises, the real money supply falls. The fall in the real money supply creates a shortage of money at the prevailing interest rate, so the interest rate rises. As with the other approach, the higher interest rate reduces consumption and desired investment and therefore reduces the real quantity of goods and services demanded.

There are two versions of the foreign trade effect. The first approach is direct: As the domestic price level rises, exports fall and imports rise as buyers substitute foreign goods for the more expensive domestic goods. The second approach, often called the *exchange rate effect,* follows from the interest rate effect. As the price level rises, the interest rate rises. A higher interest rate makes domestic bonds more attractive. As a result, domestic residents will want to buy fewer foreign bonds, and foreign residents will want to buy more domestic bonds. In the market for foreign exchange, that translates into a decrease in the supply of the domestic currency and an increase in the demand for domestic currency. That will cause the value of the currency (the exchange rate) to rise. A rise in the exchange rate means that domestic goods become more expensive relative to foreign goods. Exports will fall and imports will rise. In other words, the increase in the price level will reduce net exports, and so the real quantity of output demanded falls.

Be aware that the aggregate demand curve is not just the summation of all market demand curves, and the reasons for their negative slopes differ. For a market demand curve, a change in price is considered holding all other prices constant. It therefore represents a change in *relative* price. A market demand curve is negatively sloped because a relative price change creates income and substitution effects. A higher price causes people to substitute away from the good, and it reduces their real income. For a normal good, lower real income reduces the quantity of the good purchased. For the aggregate demand curve, a price change represents an *absolute* change in the overall price level. There can be no income effect for the macroeconomy as a whole because when the overall price level rises, both the prices people pay and the prices people receive increase. For every buyer there is a seller, so one person's higher expense is another person's higher income. To the extent that people buy more foreign goods and fewer domestic goods as the domestic price level rises, there can be a type of macroeconomic substitution effect. But as we shall see, this effect is likely to be small.

At the intermediate level, the interest rate effect is emphasized, though the other two are sometimes mentioned. In intermediate courses, the aggregate demand curve is typically derived from an IS/LM model. The interest rate effect is clearly illustrated in the IS/LM framework, as a higher price level shifts the LM curve left (upward). The result is a higher interest rate and lower real GDP in equilibrium. To the extent that the higher interest rate drives up the exchange rate, the IS curve will shift left (downward) as net exports fall. The wealth effect of a higher price level reduces consumption and so also shifts the IS curve to the left (downward). In all three cases, a higher price level is associated with lower equilibrium real GDP.

There is an alternative derivation of the aggregate demand curve based on the quantity equation, MV = PY, where M is the money supply, V is the income velocity of money, P is the price level, and Y is real GDP. MV is total spending on final goods and services. For a given value of MV, P and Y must be inversely related. In this formulation, the aggregate demand curve is a rectangular hyperbola. This version of the aggregate demand curve is less popular and is used primarily to show how changes in the money supply affect aggregate demand (Mishkin, 2007). It is less useful for showing how individual components of spending affect aggregate demand. That is because one would have to explain how a change in investment, for example, affects velocity.

## The Aggregate Supply Curve

There are two distinct textbook approaches to aggregate supply. In the simpler version, the aggregate supply curve is horizontal at the current price level for all values of output below full employment, and vertical at full employment. The idea is that as long as there is unemployment, increases in aggregate demand will increase output, not prices. Once the economy reaches full employment, further increases in demand can only drive up prices. A slightly more sophisticated version of this approach has a positively sloped region as the economy nears full employment (see Figure 32.1). The idea is that some industries will reach full employment before others, so as demand rises, some industries will respond by increasing output, while other industries will be unable to produce more and so will raise prices.

The second approach is more sophisticated and leans heavily on the research about the Phillips curve done by Nobel laureates Edmund Phelps (1967) and Milton Friedman (1968). Phelps and Friedman argued that there was a short-run trade-off between inflation and unemployment, but that the trade-off vanished in the long run. In other words, a higher inflation rate may allow the economy to move above its full-employment level of output temporarily, but not permanently.

The aggregate supply version of the Phillips curve analysis posits a long-run aggregate supply curve that is vertical at the full-employment level of output (see Figure 32.2). This corresponds to the economy's "natural" rate of unemployment that allows for only frictional and structural unemployment, but no cyclical unemployment. It is vertical because, in the long run, real GDP is determined by real forces such as the availability of resources and technology, not by the price level. Over time, the long-run aggregate supply curve will shift right as technology improves and as the economy acquires more capital and labor. Its position is also affected by tax rates and other things that affect the incentives to supply capital and labor, create businesses, and develop new technology.

Deviations from full employment can occur when the actual price level deviates from the expected price level. There is no consensus as to why this is the case, but three different explanations are commonly presented (Mankiw, 2007).

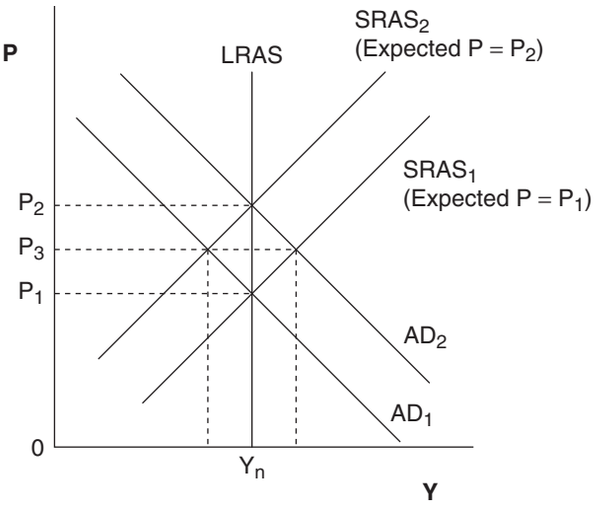


**Figure 32.2** The Enhanced AD/AS Model

SOURCE: Adapted from Mankiw (2007, p. 762).

One explanation, called the *sticky-wage model,* is based on the idea that wages are not perfectly flexible and may be fixed in the short run. Wages are set based on some expected price level, so if prices turn out to be higher than expected, real wages will be lower than expected. In response, firms will hire more workers and increase output. In other words, the aggregate supply curve will have a positive slope in the short run. The effect will not persist, however, because nominal wages will eventually adjust to reflect the higher level of prices. When price expectations are adjusted upward, workers will demand higher wages. Employment will then return to its long-run value. In terms of the model, the short-run aggregate supply curve will shift left (see Figure 32.2).

Another explanation for the short-run aggregate supply curve's positive slope begins with the observation that many prices are sticky. That means that some firms change their prices only periodically. The prices of raw commodities like grain and metal may fluctuate daily, or even hourly, but prices of finished goods behave differently. B. Bhaskara Rao (1993) found that over 85% of U.S. transactions on final goods and services take place in "sluggish price" markets. That means that sticky prices are quite common. One reason is markup pricing. Price setters often set price by adding a markup over marginal cost. Fluctuations in demand are met with changes in output, not price. A second reason is fear of competitors. Firms do not want to be the first to raise price for fear of losing customers to competitors. A third reason is menu costs: It is expensive to frequently reprint catalogs and price lists. Moreover, when firms do change their prices, they do not all do so at once. That means that as the overall price level rises, the relative price of the output produced by firms with sticky prices falls. As a result, they experience higher than normal sales. In response, the firms increase output. The result is that a rising price level is associated with rising real GDP. Once again, however, the effect is only temporary. When the firms change their prices, they will

adjust them upward in line with the overall price level. The result is that their sales will decline and output will fall. That implies that the short-run aggregate supply curve shifts to the left (see Figure 32.2).

A third explanation assumes that some firms are temporarily fooled. When the price of their output rises due to an increase in the overall price level, these firms misperceive it as an increase in the relative price of their output. In response, they increase production. As in the previous two cases, the increase in real GDP is only temporary. Firms will eventually correct their perceptions and recognize that the relative price of their output has not increased. When that happens, they will reduce production. The short-run aggregate supply curve will shift left (see Figure 32.2).

All three of these explanations can be represented in a single equation,

$$Y = Y_n + \alpha(P - P^e)$$

where Y is actual real GDP, $Y_n$ is the long-run or "natural" rate of real GDP, $\alpha$ is a positive constant, P is the actual price level, and $P^e$ is the expected price level. The equation shows that when the actual price level deviates from the price level that was expected, real GDP will deviate from its natural rate. In the long run, when expectations adjust to reality, the price level will equal the expected price level, so real GDP will equal its natural rate.

### The Adjustment to Equilibrium

As in so many other economic models, equilibrium is where the lines cross. Suppose, for example, that in Figure 32.2, $AD_2$ is the relevant aggregate demand curve and $SRAS_2$ is the relevant short-run aggregate supply curve. The economy is in both short-run and long-run equilibrium with the price level at $P_2$ and real GDP at its "natural" level, $Y_n$. Then suppose that aggregate demand falls to $AD_1$. At $P_2$, the amount of real GDP people want to buy is less than the amount firms want to sell. In the short-run, the excess supply will cause the price level to fall to $P_3$ and real GDP to fall below $Y_n$. Because the actual price level ($P_3$) is below the expected price level ($P_2$), real GDP is below its long-run value. Over time, price expectations will fall, and the short-run aggregate supply curve will shift to the right. This effect will be reinforced by the cyclical unemployment that exists when real GDP falls below $Y_n$. Unemployment puts downward pressure on wages, and that will reduce firms' costs. Eventually, the economy will return to $Y_n$ with a price level equal to $P_1$. At that point, the short-run aggregate supply curve will be at $SRAS_1$ in Figure 32.2.

## Applications and Empirical Evidence

The primary application of the model is to predict and explain short-run fluctuations in real GDP around its long-run trend. It is intended to explain how recessions and booms occur, and why the economy tends to return to its natural rate of real GDP. According to the model, deviations from the economy's long-run trend are the result of shocks to either demand or supply.

### Demand Shocks

For a given price level, a change in any of the components of aggregate demand will shift the aggregate demand curve. Recall that aggregate demand is the sum of consumption and desired investment, government purchases of goods and services, and net exports. A change in any of these can shift the aggregate demand curve. If any of them fall, aggregate demand will shift to the left. As detailed at the end of the previous section, that will cause a temporary decline in real GDP, otherwise known as a *recession*. If any of the components rise, aggregate demand will shift right. For example, in Figure 32.2, begin with $AD_1$ and $SRAS_1$. At $P_1$ and $Y_n$, the economy is in short-run and long-run equilibrium. If aggregate demand rises to $AD_2$, the short-run result is an increase in the price level to $P_3$ and an increase in real GDP above $Y_n$. The economy is in a boom period. Eventually, price expectations will be revised upward, and the short-run aggregate supply curve will shift left to $SRAS_2$. When that happens, the boom is over and the economy returns to its long-run trend.

The key point is that fluctuations in aggregate demand cause fluctuations in real GDP. The components of aggregate demand are each influenced by a large number of variables. It is therefore not surprising that aggregate demand fluctuates. The model shows how fluctuations in demand are translated into fluctuations in real GDP.

Of all of the components of aggregate demand, desired investment spending is the most volatile. As John Maynard Keynes (1936/1964) emphasized, investment is extremely sensitive to the state of expectations. That is because investors face known present costs and uncertain future returns. The decision to build a factory is an act of faith. The investor can never be certain what future demand for the factory's output will be. A recession may suppress demand and lower price. Technological change may make it obsolete. There is no way to precisely calculate the return on the factory, and Keynes argued that one typically does not even know enough to establish probabilities. It is no surprise that waves of optimism or pessimism among investors can lead to significant fluctuations in investment, aggregate demand, and real GDP.

Consumption spending is also affected by the state of expectations, though to a lesser degree. That is primarily because spending on necessities such as food, shelter, and transportation cannot be avoided. But consumption spending on some goods, such as furniture and entertainment, can be postponed. An old sofa can be made to last longer. If consumers fear a recession that might put their jobs in jeopardy, they will try to save more and consume less. If they are optimistic and believe that the economy will boom, they may expect higher income in the future. That

could lead them to save less (or borrow more) and consume more now.

Export demand depends to a significant degree on the health of our trading partners' economies. If one or more of them go into a recession, their demand for our exports will fall. That, in turn, will reduce our aggregate demand. Other things equal, the result would be to create a recession here too.

While the model correctly predicts that a decline in aggregate demand will reduce real GDP, it also predicts that the overall price level will fall. One weakness of the model becomes apparent when one realizes that the overall price level in the United States has not fallen over the course of a year since 1955. One would be hard-pressed to argue that there have not been any negative demand shocks over that period. Defenders of the model argue that there are other forces in the economy that keep the price level rising even in the face of falling aggregate demand. They contend that falling aggregate demand puts *downward pressure* on prices—that is, will slow the rate of growth of prices. That may well be true, but that is not what the model says. As some economists have argued, the model is much better at explaining output movements than price movements.

### Supply Shocks

Broadly speaking, there are two types of supply shocks: input-price shocks and technology/resource shocks. Input-price shocks lead to sudden changes in the expected price level and therefore shift the short-run aggregate supply curve. Technology/resource shocks affect the economy's production function and shift both the long-run and the short-run aggregate supply curves.

#### *Input-Price Shocks*

The aggregate demand and aggregate supply model became popular primarily because of the oil price shock that hit the West in the mid-1970s. The model was able to correctly predict the consequences of the shock. The 1973 Arab–Israeli war led Arab oil-exporting countries to embargo the West. The price of oil more than doubled in a short period of time. The price of oil is important because it contributes to the cost of producing and distributing almost everything. Inputs must be shipped to factories and output must be shipped to stores. Workers must travel to work. When the price of oil increased suddenly, people understood that costs, and therefore prices, would soon increase too. In other words, the expected price level increased, so the short-run aggregate supply curve shifted to the left.

In terms of Figure 32.2, begin with $AD_1$ and $SRAS_1$, with the economy in long-run equilibrium at $P_1$ and $Y_n$. The higher expected price level shifted the short-run aggregate supply curve to $SRAS_2$. In the short run, the price level rises to $P_3$, and real GDP falls below $Y_n$. The combination of falling output and rising prices is often called *stagflation.* The U.S. economy saw an increase in the unemployment rate from 4.9% in May of 1973 to 9.0% in May of 1975. Over the same period, the price level as measured by the Consumer Price Index increased 21.15%. A second oil-price shock hit the United States in the late 1970s, when the price of oil again more than doubled. The results were similar. The unemployment rate increased from 6.0% in May of 1978 to 7.5% in May of 1980, and the price level rose 26.69% over the same period.

In the mid-1980s, the price of oil moved sharply in the other direction. From 1982 to 1986, the price of crude oil fell more than 50%. The drop in the price of oil reduced the expected price level and shifted the short-run aggregate supply curve to the right. In terms of Figure 32.2, begin with $AD_2$ and $SRAS_2$, with the economy in long-run equilibrium at $P_2$ and $Y_n$. The fall in the expected price level shifts the short-run aggregate supply curve to $SRAS_1$. The model predicts that the price level will fall to $P_3$ and real GDP will rise above its natural rate. As the model predicts, the U.S. unemployment rate fell from 9.4% in May of 1982 to 7.2% in May of 1986. The price level rose 13.65% over the same period, contrary to the model's predictions. An explanation is that aggregate demand increased over the period, which would tend to drive up the price level.

#### *Technology/Resource Shocks*

The aggregate demand and aggregate supply model is designed to explain business cycles, but it is worth briefly mentioning a few long-run effects. Improvements in technology raise the productivity of a nation's resources and thereby increase the natural rate of GDP. As a result, both the long-run and short-run aggregate supply curves shift to the right. The acquisition of more resources through immigration, natural population growth, and capital accumulation will have the same effect on the two curves. This is another way of saying that in a growing economy, the natural rate of GDP will rise over time.

## Policy Implications

The AD/AS model as usually presented suggests that the economy will recover from adverse shocks on its own. Yet some economists argue that the process may take a long time. That is because the price level may be slow to adjust, especially in the downward direction. In other words, sticky prices and sticky wages may significantly delay the restoration of full employment. The model offers guidance to those who believe that policy can speed the process along.

### Policy Responses to Demand Shocks

If aggregate demand falls, real GDP will fall and unemployment will rise. Either fiscal policy or monetary policy (or both) can be used to increase aggregate demand. Fiscal policy uses the government's power to spend and tax. To boost aggregate demand, the government can increase its own spending or it can cut taxes so as to stimulate private

spending. The net effect on aggregate demand depends on the relative magnitude of two conflicting forces. One is the *multiplier effect.* An increase in spending raises the income of those who receive the new spending. They, in turn, will spend some of their new income, which will raise the income of somebody else. In this fashion, the initial increase in spending can induce additional spending elsewhere in the economy. The net result is that total spending in the economy can rise by a multiple of the initial increase.

The second effect is called *crowding out.* The increases in income set in motion by the multiplier effect will raise the demand for money. That will cause the interest rate to rise, which will reduce investment, consumption, and net exports. In this fashion, the initial fiscal policy action can "crowd out" other forms of spending. The crowding out will reduce the amount by which aggregate demand rises. The net effect on aggregate demand depends on the relative strength of the multiplier and crowding out effects.

Monetary policy can also be used to raise aggregate demand. If the central bank increases the money supply, the rate of interest will fall. That, in turn, will tend to raise investment, consumption, and net exports. Keynes argued in his *General Theory* (1936/1964) that monetary policy may be ineffective in a severe recession or depression because even very low interest rates may not raise spending sufficiently if investors and consumers have pessimistic expectations about the future.

### Policy Responses to Supply Shocks

Policy makers face a dilemma when confronted with an adverse supply shock. Suppose a price shock shifts the short-run aggregate supply curve to the left, which raises prices and reduces output and employment. If policy makers raise aggregate demand in order to restore full employment, the price level will rise even more. If policy makers lower aggregate demand in order to stabilize prices, output and employment will fall even farther. Any policy that changes aggregate demand will make either inflation or unemployment worse.

The only policy that *might* work to reduce both inflation and unemployment would be one that acts directly on the short-run aggregate supply curve. The options open to policy makers are limited. One possibility is to reduce business and payroll taxes in order to reduce business costs. To the extent that such a move would counteract the increase in costs caused by the price shock, the expected price level might be reduced. That would shift the short-run aggregate supply curve back to the right.

## Future Directions

The appeal of the AD/AS model is obvious. Its superficial similarity to the familiar market demand and supply model makes it easy to teach. That, together with the power of habits of thought and inertia, suggests that the model has a bright future. But as was mentioned in the introduction, the model has its critics.

### The Slope of the Aggregate Demand Curve

Consider the three reasons given for the negative slope of the aggregate demand curve. The wealth effect is much discussed, but the estimates are that it is trivially small. Nobel laureate James Tobin (1993) estimated that the wealth effect is so weak that a 10% drop in the price level would increase spending by only 0.06% of GNP. In other words, major changes in the price level would not have much of an effect on aggregate spending. As Bruce Greenwald and Nobel laureate Joseph Stiglitz (1993) argue, the excessive attention given to the wealth effect "hardly speaks well for the profession" (p. 36).

Another problem with the wealth effect is that it is selective about what types of wealth are affected by a price-level change. Specifically, it ignores the effect of a change in the price level on the real value of debt. A price-level decrease, for example, will increase the real value of debt. And while it is true that the real gains to creditors equal the real losses of debtors, the effect on spending is not neutral. This is for two reasons. The first is that, by definition, the propensity to spend is higher for debtors than for creditors. That means that the price-level decrease will reduce the economy's overall propensity to spend. The second reason is that a falling price level makes it harder to repay debt. One can reasonably expect, and history has amply demonstrated, that a decrease in the price level will increase the proportion of bad loans and the number of bankruptcies. Taken together, the implication is that a fall in the price level is likely to *reduce* the real value of spending, not increase it as the aggregate demand curve claims.

The foreign trade effect also has problems. The version of it that argues that a higher price level will make domestic goods more expensive and so reduce net exports is strictly true only if the exchange rate is fixed. If the exchange rate is flexible, then one would expect the rising domestic price level to reduce the value of the currency. The price increase and the currency depreciation should roughly cancel each other out, with minimal effect on net exports. The other version argues that the higher interest rate associated with a higher price level will drive up the exchange rate and so reduce net exports. Yet once again, the higher price level will itself reduce the value of the currency. The overall effect on net exports will be small.

According to Tobin (1993), the interest rate effect was first discussed by Keynes (1936/1964). Yet Keynes himself did not put much faith in it. The main reason is that changes in the price level that trigger the interest rate effect are likely to alter expectations. This illustrates a central problem for the model: In discussions of short-run aggregate supply, expectations have a central role. Yet in discussions of the aggregate demand curve, expectations are completely ignored. Consumers and investors are implicitly assumed to never change their price expectations.

Consider the reasonable possibility that an increase in the price level will cause people to expect even higher prices in the future. A normal reaction might be to buy *more* now to avoid paying even higher prices later. If that happened, the aggregate demand curve could have a *positive* slope. Alternatively, if the aggregate demand curve is drawn holding price expectations constant (as it is), then it would shift to the right when the price level increased.

The more general point is that it is not at all clear that a higher price level will reduce the quantity of goods and services people buy. At best, the aggregate demand curve is extremely steep, almost vertical. If that is the case, then why bother to put the price level on the vertical axis at all? As was suggested earlier, the model is much better at predicting changes in real GDP than it is at predicting changes in the price level. The prominent role of the price level in the model is no doubt an attempt to blend neoclassical economics (which views price adjustments as the primary equilibrating force) with Keynesian economics (which views output adjustments as the primary equilibrating force). The result is not always satisfactory.

### Incompatibility of the Aggregate Demand and Aggregate Supply Curves

The aggregate demand curve is in no sense a behavioral relationship like a market demand curve. The aggregate demand curve is a set of equilibrium points. This is seen clearly when the aggregate demand curve is derived from the IS/LM model, as it usually is in intermediate macroeconomics classes. The intersection of the IS and LM curves shows the equilibrium rate of real GDP where both the goods market and the money market are in equilibrium. As the price level changes, the LM curve shifts. By repeatedly changing the price level, one can generate an aggregate demand curve that shows the relationship between the price level and the rate of real GDP where the goods market and money market are both in equilibrium.

Implicit in this story is the assumption that production rises along with demand to keep the goods market in equilibrium (Barro, 1994; Nevile & Rao, 1996). As the price level falls and the LM curve shifts right (downward), the rate of interest falls and spending rises. This is the familiar interest rate effect. Output is assumed to be forthcoming to match the increase in spending. It is a manifestation of Hansen's law that demand creates its own supply. To put the matter differently, the derivation of the aggregate demand curve implies that firms can always sell everything they care to produce at whatever the price level happens to be.

The logical problem should now be apparent. The aggregate supply curve presents a much different relationship between the price level and the level of output that firms produce. Worse yet, the AD/AS model argues that if the price level is above its equilibrium value, then there will be an excess supply of goods. How can that be when the aggregate demand curve is derived with the assumption that the goods market is in equilibrium?

One cannot argue that the aggregate demand curve is a notional demand curve akin to a market demand curve, where the relationship between price and quantity demanded is hypothetical. Instead, the aggregate demand curve shows us what the *market clearing* level of output is at various price levels. And if the market is clearing, there cannot be an excess supply. This inconsistency is the primary reason that prominent macroeconomist Robert Barro (1994) said that the model is "unsatisfactory and should be abandoned as a teaching tool" (p. 1).

David Colander (1997) argues that the aggregate demand curve should be called the *aggregate equilibrium demand curve* so that its nature is apparent to students and professors alike. He goes so far as to say that the aggregate demand curve was deliberately named incorrectly in order to mislead students into thinking that macroeconomics was nothing more than "big micro."

Colander also argues that an aggregate supply curve may not exist for the same reason that a supply curve does not exist in an imperfectly competitive market. In such markets, output decisions cannot be separated from demand. A price-setting firm must choose price and output at the same time and cannot do so without reference to demand. Colander argues that most firms charge a markup over cost, and the amount they produce is based on expected demand. It is therefore impossible to create a supply function independent of the demand function.

The fundamental point is that in a Keynesian short-run world, output adjusts to clear markets. In a neoclassical long-run world, prices adjust to clear markets. The AD/AS model is an attempt to have it both ways. In it, price-level changes work to equate the quantity of real GDP demanded and supplied even in the short run. To do so, there must be separate supply and demand functions. But if Keynes and Colander are correct, one cannot separate production decisions from demand.

## Conclusion

In a famous article, Friedman (1953) argued that the real test of a model is how well it predicts. As noted previously, the AD/AS model became popular primarily because it correctly predicted the effects of the price shocks of the 1970s. It also does a pretty good job of predicting how shocks to the economy affect real GDP. But the model predicts that a decrease in aggregate demand or an increase in aggregate supply will reduce the price level, and that has not happened in the United States for over 50 years. Certainly aggregate demand has fallen and aggregate supply has increased in the last half century. Defenders of the model argue that what the model is really saying is that such events put downward pressure on the price level, so that the rate of inflation will decrease. Yet that begs the question as to why the model has the price level on the vertical axis instead of the inflation rate.

Models with the inflation rate on the vertical axis exist and occasionally appear in textbooks (Jones, 2008; Taylor &

Weerapana, 2009). These models typically derive the aggregate demand curve from a monetary policy rule (Taylor's rule) that describes how the central bank reacts to changes in the inflation rate. As the inflation rate rises, the central bank increases the nominal interest rate by more than the rise in the inflation rate. The effect is to increase the real interest rate, which reduces the real volume of spending. The approach avoids the problems with the aggregate demand curve discussed previously. John B. Taylor and Akila Weerapana's model has no aggregate supply curve. Instead, the microfoundations of price setting are discussed for the purpose of emphasizing how inflation is self-perpetuating in the short run. Charles Jones uses a variant of the Phillips curve as an aggregate supply curve. Critics of these models complain that they lean too heavily on a policy rule that may not always exist. Yet all models ultimately depend on specific institutions (such as property rights) and habits of thought. To look for a macroeconomic model devoid of an institutional foundation is to look for a chimera.

The fundamental problem is that the macroeconomy is a dynamic, complex thing. Static models cannot capture its nuances. A model simple enough to teach to undergraduates is bound to have flaws; in teaching, there is always a trade-off between accuracy and accessibility. The real question then becomes the extent to which the model convinces students of things that are not true.

Thomas Kuhn (1962/1996) famously observed that displacing well-entrenched theories is never easy. For one thing, the need for an alternative is not always clear to most scientists, and some actively suppress any heresies. It is safer to go along with the conventional wisdom. For a new approach to succeed, it must be clearly and demonstrably superior. If the AD/AS model is ever to be replaced, its critics must develop superior alternatives.

## References and Further Readings


Barrens, I. (1997). What went wrong with IS-LM/AS-AD—and why? *Eastern Economic Journal, 23,* 89–99.

Barro, R. J. (1994). The aggregate-supply/aggregate-demand model. *Eastern Economic Journal, 20,* 1–6.

Brownlee, O. H. (1950). The theory of employment and stabilization policy. *Journal of Political Economy, 58,* 412–424.

Colander, D. (1995). The stories we tell: A reconsideration of the AS/AD analysis. *Journal of Economic Perspectives, 9,* 169–188.

Colander, D. (1997). Truth in labeling, AS/AD analysis, and pedagogy. *Eastern Economic Journal, 23,* 477–482.

Dutt, A. K. (1997). On the alleged inconsistency in aggregate supply/aggregate demand analysis. *Eastern Economic Journal, 23,* 469–476.

Dutt, A. K. (2002). Aggregate demand-aggregate supply analysis: A history. *History of Political Economy, 34,* 322–363.

Dutt, A. K., & Skott, P. (1996). Keynesian theory and the aggregate supply/aggregate-demand framework: A defense. *Eastern Economic Journal, 22,* 313–331.

Fields, T. W., & Hart, W. R. (2003). Bridging the gap between interest rate and price level approaches in the AD-AS model: The role of the loanable funds market. *Eastern Economic Journal, 29,* 377–390.

Friedman, M. (1953). The methodology of positive economics. In *Essays in positive economics* (pp. 3–43). Chicago: University of Chicago Press.

Friedman. M. (1968). The role of monetary policy. *American Economic Review, 58,* 1–17.

Greenwald, B., & Stiglitz, J. (1993). New and old Keynesians. *Journal of Economic Perspectives, 7,* 23–44.

Hansen, R. B., McCormick, K., & Rives, J. M. (1985). The aggregate demand curve and its proper interpretation. *Journal of Economic Education, 16,* 287–296.

Hansen, R. B., McCormick, K., & Rives, J. M. (1987). The aggregate demand curve: A reply. *Journal of Economic Education, 18,* 47–50.

Jones, C. I. (2008). *Macroeconomics.* New York: W. W. Norton.

Keynes, J. M. (1964). *The general theory of employment, interest and money.* New York: Harcourt, Brace & World. (Original work published 1936)

Kuhn, T. S. (1996). *The structure of scientific revolutions.* Chicago: University of Chicago Press. (Original work published 1962)

Mankiw, N. G. (2007). *Principles of economics* (4th ed.). Mason, OH: Thomson-Southwestern.

McCormick, K., & Rives, J. M. (2002). Aggregate demand in principles textbooks. In B. B. Rao (Ed.), *Aggregate demand and supply: A critique of orthodox macroeconomic modelling* (pp. 11–23). London: Macmillan.

Mishkin, F. S. (2007). *The economics of money, banking, and financial markets* (8th ed.). New York: Pearson-Addison Wesley.

Nevile, J. W., & Rao, B. B. (1996). The use and abuse of aggregate demand and supply functions. *The Manchester School, 64,* 189–207.

Palley, T. (1997). Keynesian theory and AS/AD analysis: Further observations. *Eastern Economic Journal, 23,* 459–468.

Phelps, E. S. (1967). Phillips curves, expectations of inflation and optimal unemployment over time. *Economica, 34,* 254–281.

Rabin, A. A., & Birch, D. (1982). A clarification of the IS curve and the aggregate demand curve. *Journal of Macroeconomics, 4,* 233–238.

Rao, B. B. (1993). The nature of transactions in the U.S. aggregate goods market. *Economics Letters, 41,* 385–390.

Rao, B. B. (Ed.). (1998). *Aggregate demand and supply: A critique of orthodox macroeconomic modelling.* London: Macmillan.

Rao, B. B. (2007). The nature of the ADAS model based on the ISLM model. *Cambridge Journal of Economics, 31,* 413–422.

Reynolds, L. G. (1988). *Economics* (6th ed.). Homewood, IL: Irwin.

Saltz, I., Cantrell, P., & Horton, J. (2002). Does the aggregate demand curve suffer from the fallacy of composition? *The American Economist, 46,* 61–65.

Taylor, J. B., & Weerapana, A. (2009). *Economics* (6th ed.). New York: Houghton Mifflin.

Tobin, J. (1993). Price flexibility and output stability: An old Keynesian view. *Journal of Economic Perspectives, 7,* 45–65.

Truet, L. J., & Truet, D. B. (1998). The aggregate demand/supply model: A premature requiem? *The American Economist, 42,* 71–75.

# 33

# IS-LM Model

Fadhel Kaboub
*Denison University*

The birth of modern macroeconomics is often credited to John Maynard Keynes (1881–1946) and his classic 1936 book *The General Theory of Employment, Interest and Money.* Keynes's agenda in 1936 was twofold. First, he wanted to save the discipline of economics from being completely dismissed and ridiculed by policy makers and by the general public who were fed up with the neoclassical laissez-faire policy advice that dominated the Great Depression era of the 1920s and the 1930s. Second, Keynes feared the rapid political shift to the far left in Western Europe and the growing sympathy for and admiration of the ideals promoted by communism, not just by the working class but by a certain segment of the bourgeoisie. As a matter of fact, Keynes was rushing to publish his *General Theory* in order to provide a plausible explanation for the Great Depression and lay out an effective policy response to bring capitalist societies back on the road to economic prosperity. In seeking to achieve these two goals, Keynes had to completely divorce himself from the neoclassical theoretical apparatus that had paralyzed the economics profession during the Great Depression. The ensuing Keynesian revolution has brought forward (a) the theory of effective demand, (b) the concept of quantity adjustment rather than price adjustment, and (c) an explicit treatment of the time-dependent concepts (uncertainty about the future, speculation, "animal spirits").

Keynes's work sent a shockwave through the economics profession and was met with a lot of resistance by the gatekeepers of the economics discipline. Some younger economists, however, received Keynes's work with a lot of enthusiasm and support, but many of them could not help but interpret it by using the neoclassical theoretical framework that had been engraved into their brains. Therefore, the Keynesian revolution was quickly reincorporated into the neoclassical framework that Keynes sought to destroy.

In 1937, just a little over a year after the publication of the *General Theory,* Sir John H. Hicks (1904–1989) published one of his most famous articles, titled "Mr. Keynes and the 'Classics': A Suggested Interpretation." This article was first presented in September 1936, as an attempt to explain Keynes's theory to econometricians and mathematical economists. The outcome of this article turned out to be the dominant macroeconomic model until the mid-1960s: the IS-LL model, which later became known as the IS-LM model. Hicks's interpretation of Keynes became a classic textbook section in all macroeconomics textbooks and an essential tool for policy analysis for several decades. The model, however, was far from being faithful to Keynes's economic analysis, and it soon came under harsh scrutiny by post-Keynesian economists. By 1980, just 8 years after winning the Nobel Prize in economics, Hicks admitted that there were major flaws in his model and accepted the post-Keynesian critique. The IS-LM model soon lost its premiere position in academic research but remained a classic textbook exercise in intellectual gymnastics and an easy tool to communicate policy advice to policy makers.

This chapter aims at presenting the history of the development of the IS-LM model, its theoretical formulation and policy applications, the critiques of the model, and the direction taken by the economics profession after the demise of the IS-LM model.

## The Essence of Keynes's *General Theory*

To be able to understand the development of the IS-LM theory, one has to understand the basics of the revolutionary theory that Keynes introduced to the discipline of economics in 1936. Only then can one understand the context in which

Hicks worked to develop the IS-LM model. For Keynes, economic growth is driven by effective demand, investment is driven by expectations about future profits, and money is not neutral.

Unlike neoclassical economics where economic agents operate in notional or theoretical time, in the Keynesian model, economic agents operate in real historical time (minutes, hours, days, years, etc.) under conditions of uncertainty in which the future is unknowable and the past is unchangeable. Faced with such an environment, individuals make arbitrage decisions with regard to which assets they wish to hold over time. Each asset gets a return composed of four components: $q - c + l + a$, where $q$ is the expected yield, $c$ is the carrying cost, $l$ is the liquidity premium, and $a$ is the appreciation or depreciation.

At equilibrium, all assets earn the same expected return. If an asset has a demand price higher than its supply price, then firms will produce more of it, but as its production increases, its rate of return will fall and becomes equal to the rates of return of all other assets. When consumers and firms are optimistic about the future and feel confident about their financial situation, the expected returns on capital equipment rise above the expected return on money (i.e., the interest rate), which leads to an increase in investment, thus boosting output and employment. Conversely, when the economy is overtaken by pessimistic expectations, consumers and firms prefer to remain liquid, thus abstaining from spending on consumption and investment goods, which leads to a rise in unemployment. The business cycle is therefore driven by these waves of optimistic and pessimistic expectations.

According to Keynes, unemployment is due to money's very specific nature as the most liquid asset in the economy with a near-zero elasticity of production, small elasticity of substitution, and no carrying cost. In other words, when people want to hold more money (liquidity), no significant amount of labor is directed to producing it, unlike capital equipment, which, when it is in high demand, requires additional labor input to produce it. Therefore, Keynes's conclusion was that an environment must be created that is conducive to more investment and less hoarding of money. Hence his three policy recommendations: (1) "parting with liquidity" (giving up liquid assets in exchange for employment-creating illiquid assets), (2) "euthanizing the rentiers" by lowering the interest rate so much that nobody will find it profitable to save money (because expected returns on money are less attractive than expected returns on capital), and (3) "socializing investment" through the creation of a new kind of capitalist culture of cooperation between private and public authorities (Keynes, 1936, chap. 24).

One would expect that an accurate interpretation of Keynes's theory would remain faithful to the above basic principles. Joan Robinson argued that Keynes "was himself partly to blame for the perversion of his ideas" (Robinson, 1962, p. 90) and that he "himself began the reconstruction of the orthodox scheme that he had shattered" (Robinson, 1971, p. ix). Robinson was referring to Keynes's last chapter in the *General Theory,* where he argued that "if our central controls succeed in establishing an aggregate volume of output corresponding to full employment as nearly as practicable, the classical theory comes into its own again from this point onwards" (Keynes, 1936, p. 378). This statement in the *General Theory* led many to believe that Keynes was admitting the validity of neoclassical theory and its general applicability. However, what Keynes meant was that once full employment is achieved through government spending, the neoclassical "special case" would be valid because neoclassical theory assumes that the economy operates at full employment. Hicks's interpretation of Keynes's work was capable of incorporating some of Keynes's conclusions in a very clever way but still managed to validate some of the basic neoclassical principles. That is what led to the creation of the so-called neoclassical-Keynesian synthesis school of thought.

## Development of the IS-LM Model

When Keynes wrote the *General Theory* in 1936 he was launching an attack against the "classics," by which he did not mean the classical political economy of Adam Smith, William Petty, James Mill, David Ricardo, Thomas Malthus, and Karl Marx but rather the neoclassical school of thought of John Stewart Mill, Leon Walras, Carl Menger, William Stanley Jevons, Alfred Marshall, Francis Y. Edgeworth, and Arthur C. Pigou. Keynes's contribution recognized that capitalist economies do not have any natural tendencies to stabilize at full-employment equilibrium, but instead the system is driven by destabilizing forces that push the economy into a downward spiral if the government does not intervene to alleviate unemployment.

According to Keynes, economic recessions and depressions are due to the lack of effective demand. His analysis highlighted the importance of uncertainty and expectations to the determination of the level of economic activity. In real historical time (minutes, days, and years, as opposed to notional or analytical time), the past is unchangeable and the future is unknowable; therefore, both consumption and business investment decisions are made under conditions of uncertainty. Spending decisions always depend on future expectations. A wave of positive expectations about the future will lead to an increase in business investment, which increases output, income, and employment. For Keynes, the economy moves from one equilibrium point to another through a process of quantity adjustment rather than price adjustment. The neoclassical price mechanism (prices, wages, and interest rates) plays no role in the Keynesian system. Prices are administered through relative market power relations rather than perfectly competitive market conditions. Furthermore, Keynes argued that recessions are not to be blamed on the lack of price flexibility.

If anything, price rigidity helps prevent an acceleration of the downward spiral during bad economic times.

Keynes's approval of Hicks's interpretation of his work was primarily based on the fact that the IS-LM model was able to capture the importance of effective demand, the possibility of underemployment equilibrium, and the important role of fiscal policy to achieve and maintain full employment. That was sufficient for Keynes to publicly recognize Hicks's work as a step in the right direction, even though he criticized the IS-LM model in private communications and thought that it was too simplistic to capture the complexities of economic reality.

The IS-LM model starts with Keynes's rejection of the "classical dichotomy" theory that supports the neutrality of money. Keynes argued that changes in the quantity of money will have a real impact on the quantity of output, income, and employment. There is an inevitable interaction between the monetary and the real spheres. Therefore, Hicks (1937) argued that it is necessary to solve for the money and real markets simultaneously. The IS-LM model is essentially the superposition of two curves: the LM curve, which is derived from the money market, and the IS curve, which is derived from the goods market.

The equilibrium condition in the goods market is $S = I$ (saving = investment), and the equilibrium condition in the money market is $L = M$ (money demand = money supply). The basic equations of the IS-LM model are defined as follows:

$$S = a + bY + c\,i. \tag{1}$$

$$I = d + eY + f\,i. \tag{2}$$

$$L = A + BY + C\,i. \tag{3}$$

$$M = M_0. \tag{4}$$

*a:* autonomous savings ($a < 0$)
*b:* marginal propensity to save ($b > 0$)
*c:* interest elasticity of savings ($c > 0$)
*d:* autonomous investment ($d > 0$)
*e:* marginal propensity to invest ($e > 0$)
*f:* interest elasticity of investment ($f < 0$)
*A:* autonomous demand for money ($A > 0$)
*B:* transactions demand for money ($B > 0$)
*C:* speculative demand for money ($C < 0$)
*M:* exogenous money supply determined by the central bank
*Y:* aggregate output and income
*i:* interest rate

To solve for the IS-LM equations, we set $I = S$ and $L = M$, and then we solve for the IS and LM equations. The resulting equations are the following:

$$i_{IS} = \frac{(d-a)}{(c-f)} + \frac{(e-b)}{(c-f)} \bullet Y. \tag{5}$$

$$i_{LM} = \frac{(M-A)}{C-(B/C)} \bullet Y. \tag{6}$$

Finally, to solve for the equilibrium level of interest rate ($i^*$) and the equilibrium level of output and income ($Y^*$), we have to set $i_{IS} = i_{LM}$ and solve for $Y^*$, then plug the $Y^*$ value into either Equation 5 or Equation 6 to find the $i^*$ value.

$$i^* = \frac{(M-A)(b-e) - B(d-a)}{B(f-c) + C(b-e)}. \tag{7}$$

$$Y^* = \frac{(M-A)(c-f) - C(d-a)}{B(c-f) + C(e-b)}. \tag{8}$$

Equations 5 and 6 allow us to plot a downward-sloping IS curve and an upward-sloping LM curve. However, the downward slope of the IS curve is possible only under the condition that the marginal propensity to save is greater than the marginal propensity to invest ($b > e$). Assuming that $b$ is greater than $e$, the resulting IS-LM model gives us the equilibrium level of interest rate ($i^*$) and the equilibrium level of output and income ($Y^*$). This equilibrium solution must have four characteristics:

1. it must exist (i.e., the IS and LM curves must intersect),
2. it must be unique (i.e., the IS and LM curves must intersect only once),
3. it must be positive ($i^* > 0$ and $Y^* > 0$), and
4. it must be stable (i.e., any shock to the system produces temporary disequilibrium and an eventual return to the equilibrium point).

In the goods market, the investment demand function is negatively related to the interest rate. A fall in the interest rate will result in an increase in the demand for investment, leading to an increase in the level of output (income) and employment through the multiplier effect. The IS curve is steep when the investment demand function is interest inelastic, and it is flat when the investment demand function is interest elastic. It is noteworthy here to mention that the full employment level $Y_f$ is generally higher than the equilibrium level of output $Y^*$. Contrary to the neoclassical model, there are no inherent mechanisms to ensure that the equilibrium level of output would coincide with the full employment level. Therefore, under normal conditions, the economy will be sustained at the underemployment equilibrium.

The IS curve represents the equilibrium locus that captures the relationship between the interest rate and output levels. As the interest rate rises, investment falls and so does disposable income, and thus the equilibrium level of output $Y^*$ declines. Therefore, the IS curve is downward sloping: High interest rates are associated with low-equilibrium output $Y^*$, while low interest rates are associated with high $Y^*$. This is, in fact, an equilibrium locus and not a curve, which means that at any point on the IS curve, the goods market clears (neither excess supply nor excess demand). The multiplier dynamic ensures that at any point not belonging to the IS curve ($Y > Y^*$ or $Y < Y^*$), there will be an automatic return to the equilibrium locus. Thus, points to the left of the IS curve represent situations of excess demand for goods, whereas points to the right of the IS curve represent situations of excess supply of goods.

What Keynes meant by using the term *General Theory* in the title of his book is that neoclassical economics is a "special case" in which the economy settles at a full employment equilibrium, whereas his theory explains the "general case" where the economy could be at equilibrium at any level of employment, but with a general and most likely situation of unemployment—hence the concept of "underemployment equilibrium." In trying to make this argument, Keynes had to launch a direct attack on neoclassical economics and its leading advocates, including his own professor, Arthur C. Pigou (1877–1959). According to Hicks, Keynes was wrong in claiming that the classics had no theory of money, wages, and employment. In other words, Keynes rejected the classical dichotomy between the real and monetary spheres. Hicks went even further to argue that Keynes's theory and the classical theory are both special cases. The Keynesian theory could be represented by a horizontal LM curve, in which case an increase in aggregate demand would shift the IS curve to the right, thus leading to an increase in output and employment without increasing interest rates or prices. The classical model, however, could be represented by a vertical LM curve at the full employment level, in which case any increase in aggregate demand would only lead to inflation and higher interest rates because the economy is already operating at the full employment level.

Hicks's IS-LM model generalized Keynes's *General Theory* and argued that the classical model is a special case of an economy operating at full employment, but also Keynes's theory is a special case of a Great Depression economy. The IS-LM model, however, is a more general theory as it shows the operation of the economy under normal circumstances (neither a depression nor full employment) with an upward-sloping LM curve, in which case an increase in aggregate demand would lead to a simultaneous increase in output and employment, as well as interest rates.

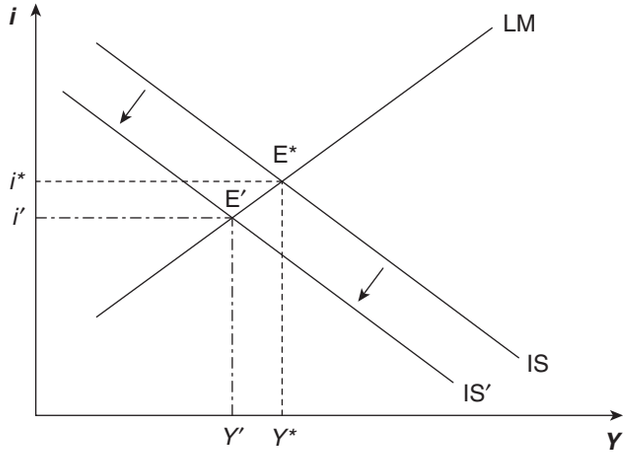


**Figure 33.1** The IS-LM Diagram

## IS-LM Policy Analysis

A major reason for the popularity of the IS-LM model from the late 1940s until the mid-1970s was its easy and convenient application in the study of alternative policy scenarios. For instance, one can easily find the impact of an increase in consumer thriftiness on the overall level of income, output, and employment. Following the IS-LM mechanism, a sudden increase in consumer thriftiness, meaning an increase in the marginal propensity to save *b*, results in a decrease in the production level *Y*, leading to a decrease in the employment level. Furthermore, as a result of the interaction of the Keynesian multiplier and the accelerator, an increase of the intended saving supposedly may result in a decrease of the actual saving. This phenomenon is known as the paradox of thrift.

Many years after the introduction of Hicks's IS-LM model, which was the first misinterpretation of Keynes's *General Theory,* Franco Modigliani (1944) revised the standard model in his PhD dissertation written under Jacob Marschak at the New School for Social Research. He proposed to add to Hicks's model the missing labor market and production function equations. As a matter of fact, Modigliani imposed the labor market-clearing assumption when he defined the labor supply function as a function of the real wage. He showed that when this assumption is added, the IS-LM model results in a price elasticity of money supply equal to 1. He concluded that money is *completely neutral,* as neither the interest rate nor the employment and output levels are affected by an increase in the money supply. In short, he established that the neoclassical results can be derived from a seemingly Keynesian-like set of equations. Modigliani's results were expected because he assumed that the labor market would reach full employment, an assumption that goes against the essence of Keynes's analysis.

In the next step of his analysis, Modigliani ran his model again without the market-clearing assumption and with the assumption of rigid money wages. His results showed that the price elasticity of the money supply is less than 1. Furthermore, an increase in the money supply results in a decrease in the interest rate and an increase in both employment and output levels to the full employment level. Therefore, he concluded that money is *not neutral* and that Keynes's conclusions are now restored. Modigliani's conclusions were adopted during the 1950s and early 1960s by the conventional wisdom of the economics discipline, which is essentially a "synthesis" of neoclassical and Keynesian theory, where the results of the model in a "perfectly working" IS-LM model (i.e., in the long run) are neoclassical (i.e., full-employment equilibrium). In an "imperfectly working" IS-LM model (i.e., in the short run, money wages are rigid), however, the Keynesian conclusions are restored.

Robert W. Clower (1965) and his student Axel Leijonhufvud (1967) launched a significant attack against

the standard IS-LM presentation of Keynes's theory and started the so-called disequilibrium[1] Keynesianism movement. Both Clower and Leijonhufvud believed that the disequilibrium situation (i.e., unemployment and effective demand failure) is the result of information and coordination deficiencies.

The IS-LM model, according to Leijonhufvud, has also a fundamental weakness, causing a terrible misinterpretation of Keynes's *General Theory*. Both Keynes's model and the IS-LM model have five goods in the system (consumer goods, capital goods, labor, money, and bonds). To solve for the three relative prices (i.e., the overall price $p$, the real wage $w$, and the real interest rate $i$), the models assimilate one good into another. The IS-LM model combines consumer goods with capital goods (commodities), whereas Keynes's model, according to Leijonhufvud, combines bonds with capital goods (nonmoney assets). This is a major difference between Keynes and the IS-LM Keynesians in the aggregative structure of their models.

In the neoclassical approach, there is only one kind of output and one price for it. However, in Keynes's model, there are two kinds of output: consumption goods and investment goods. Along with the two kinds of output come two completely different price systems: a price system for current output and a price system for capital assets, the price of which is crucial in the determination of the level of investment that will take place. The interplay between these two price systems determines the level of investment and, through the multiplier, determines the level of output and employment. Unemployment results when the two price systems fail to generate a set of prices that is consistent with the full employment level of production (i.e., either the demand price of capital assets is too low or the supply price of capital assets is too high).

The more headway the IS-LM made into economics textbooks and economic policy circles, the more scrutiny and criticism it drew from the followers of Keynes who tried to remain faithful to his theoretical analysis and to vindicate his theory from all the misuse and abuse brought by the so-called IS-LM textbook Keynesians, or "bastard Keynesians," a term used by Joan Robinson to refer to economists who do not know whether the father of their theories is Keynes or Marshall. The following is a summary of the main points of criticism made by post-Keynesians such as Joan Robinson, Richard Khan, Hyman Minsky, Paul Davidson, and Sidney Weintraub, to mention a few.

The downward slope of the IS curve is possible only under the condition that the marginal propensity to save is greater than the marginal propensity to invest ($b > e$). The most reasonable assumption, however, would suggest the exact opposite. That is to say that for most industrialized economies, the marginal propensity to invest should be greater than the marginal propensity to save, especially under a very sophisticated banking system. An upward-sloping IS curve could result in the violation of some of the equilibrium conditions. An upward-sloping IS curve may not intersect with the LM curve in the first quadrant, or it may intersect with the LM curve twice if there is a liquidity trap (i.e., in the horizontal section of the LM curve). This would violate the existence and the uniqueness conditions of the equilibrium solution in the IS-LM model.

In closing his 1937 article, Hicks recognizes some limits to his general theory—namely, the fact that most variables would remain indeterminate as long as income and distribution are unknown. He also admitted that depreciation and the timing of processes are neglected in his analysis as well. By 1976, Hicks grew more dissatisfied with the IS-LM model. In his seminal article "Some Questions of Time in Economics" (1976), he recognized the importance of the irreversibility of time. He argued that Keynes's theory was a hybrid one that is divided in two parts, one "in time" and another "out of time." Keynes's theory of the marginal efficiency of capital and his liquidity preference theory are "in time" because they are forward-looking concepts that integrate the irreversibility of time and fundamental uncertainty about the future. Keynes's multiplier theory, as well as the theory of production and prices that it implies, however, is a theory "out of time." It is a theory that runs in terms of supply-and-demand curves like the old tools of neoclassical equilibrium analysis. A state of equilibrium is by definition a state in which nothing relevant changes, and hence time is put aside. Hicks conceded that his IS-LM model has reduced Keynes's *General Theory* to equilibrium economics, even though Keynes himself did not wholly disapprove of the original IS-LM interpretation.

In another seminal article published in 1980–1981, Hicks tried to distance himself from his original 1937 theory and provided a very candid assessment of the IS-LM model. He argued that there are great similarities between his work and that of Keynes prior to the publication of the *General Theory*. They have both worked on the behavior of the economy "during a period that had a past" (i.e., in real historical time). Hicks also argued that there are some differences that he did not explicitly mention in his earlier articles. First, the IS-LM model is a flexible price model with perfect competition, while Keynes's theory is a fixed price model that is consistent with unemployment. Second, both models are in the short run, but Hicks's model is an ultra-short-run model ("one week"), while Keynes's is a "one-year" model.

More than four decades after the birth of the IS-LM model, Hicks brought two major critiques of his own work in another showcase of his intellectual honesty. Hicks admitted that the absence of the labor market and the lack of dynamics in the IS-LM model constitute two major weaknesses in his theory. He explained that in Keynes's model, there is a possibility of unemployment even when all markets are in equilibrium. He also argued that even though the labor market does not exist in the IS-LM model, a labor market that is in disequilibrium can be

added to the model without leading to inconsistencies with the Walrasian framework.

Regarding the concept of time, Hicks argued that time has not really been entirely neglected in the IS-LM model, but it is rather the flow of time within a period that has been ignored. He also recognized the problems caused by superimposing a stock (in the LM curve) and a flow (in the IS curve) where IS-LM draws an instantaneous equilibrium between a stock at time *t* and a flow at a period of time. The solution to this problem according to Hicks would be to maintain a stock equilibrium over the period that implies a flow of equilibrium over the period. That is to say that stocks rolling over time would be assimilated to a flow. As a result, Hicks found himself in a deterministic model where there is no place for uncertainty and therefore no room for a liquidity preference theory. To fix this problem, Hicks invented a pseudo-deterministic model where the variables would fall within a particular range. Thus, the equilibrium would fall within the expected range.

In the end, Hicks concluded that the IS-LM model should not be such an important policy tool but rather simply a model that will explain economic conditions in a static way, knowing that the future is unlikely to be well predicted.

Despite all the weaknesses highlighted above, and despite the fact that all economists today recognize and accept the nonvalidity of the IS-LM model, we still observe the presence of the IS-LM model as a standard section in every single intermediate macroeconomics class taught in the United States. Students are often confused as to why they are asked to study a model that has been proven to be theoretically flawed and is considered useless for policy analysis. Most economists still teach the IS-LM model because of the following three reasons.

The model serves as a good pedagogical tool to improve the students' analytical skills and graphical analysis, preparing them for the more advanced mathematical models that they will encounter in graduate school.

Most undergraduate economics instructors teach the IS-LM model, so it is considered unfair to students not to familiarize them with it in case they decide to pursue a graduate program in economics where they would be expected to be familiar with the model.

The IS-LM model has been taught as a standard model for decades, which means that nearly all policy makers in Washington, D.C., have become familiar with it over the years as a standard tool that economists use to communicate their policy advice. Economists today use mathematical models that have nothing to do with the IS-LM analysis, but they still use a simple IS-LM presentation to communicate their results and policy recommendations to policy makers who lack the sophisticated training that it takes to be able to understand the new and more advanced mathematical models.

Despite the general consensus among economists about the demise of the IS-LM model, it is still common to find no reference to any critiques of the model in undergraduate textbooks or class discussions. Students typically find out about the model's weaknesses when they take a graduate course in macroeconomics or when they take a history of economic thought course at the undergraduate or graduate level.

## Conclusion

The IS-LM model developed by Hicks (1937) and later popularized by Paul Samuelson (1948) and Alvin Hansen (1953) has provided an extremely useful diagram that explains how the economy operates and how it could be managed to reach the full-employment equilibrium through the different fiscal and monetary policy adjustments. There is no doubt that the IS-LM model is the most popular macroeconomic model used in textbooks because it was extremely efficient as a pedagogical tool.

Like most mathematical models, however, the IS-LM model should be taken with a fistful of salt. It is a simple model that superimposes equilibrium in the goods market with equilibrium in the money market to produce a general macroeconomic equilibrium for the economy as a whole, with an equilibrium interest rate and an equilibrium level of aggregate output and employment. The model was able to explain the possibility of unemployment equilibrium, which was one of Keynes's most important arguments in the *General Theory*. This was sufficient for Keynes to approve of Hicks's work and to write him a letter saying that he "found it very interesting and [that he] really [has] next to nothing to say by way of criticism" (Keynes, 1937). Keynes of course knew the weaknesses of the IS-LM model and how far it stood from his *General Theory,* but he was willing to accept it as a strategy to make the economics profession accept the fact that markets do not self-adjust and that there is a necessity for government intervention to achieve full employment.

Keynes's followers (and eventually Hicks himself) laid out the key criticisms that brought the IS-LM hegemony to an end. The model conflates short- and long-term interest rates, conflates stocks with flows, ignores the importance of uncertainty and investor's expectations in determining business cycle fluctuations, and reintegrates the price mechanism back into a seemingly Keynesian model. The economics profession has moved far beyond the limits of the IS-LM model since the 1980s, but the traditional IS-LM way of thinking about the economy has remained a habit of thought in every economist's mind to this day, not necessarily in mathematical form but at least in spirit.

## Note

1. Clower and Leijonhufvud argued that Keynes had a disequilibrium approach instead of an equilibrium approach. When Keynes used the term *equilibrium,* he did not mean the same thing as the neoclassical approach, where equilibrium means that prices adjust so that all markets clear. When Keynes used the term *equilibrium,* he meant that there are no forces to cause further movements (markets might not clear). Thus, in Keynes's

theory, unemployment is an "equilibrium" in the sense that there are no forces to cause employment to increase. Clower and Leijonhufvud thought that this is disequilibrium because they used it in the way that the neoclassicals used it.

*Author's Note:* The author thanks L. Randall Wray for comments on an earlier draft of this chapter and one anonymous referee for helpful suggestions. Rana Odeh provided invaluable editorial assistance.

## References and Further Readings

Barens, I. (1999). From Keynes to Hicks—an aberration? IS-LM and the analytical nucleus of the *General Theory*. In P. Howitt, E. De Antoni, A. Leijonhufvud, & R. W. Clower (Eds.), *Money, markets and method: Essays in honour of Robert W. Clower.* Cheltenham, UK: Edward Elgar.

Chick, V. (1983). *Macroeconomics after Keynes.* Oxford, UK: Philip Allan.

Coddington, A. (1967). Keynesian economics: The search for first principles. *Journal of Economic Literature, 14,* 1258–1273.

Clower, R.W. (1965). The Keynesian counterrevolution: A theoretical appraisal. In F. Hahn & F. Breckling (Eds.), *Theory of interest rates.* London: Macmillan.

Clower, R. W., & Leijonhufvud, A. (1975). The co-ordination of economic activities: A Keynesian perspective. *American Economic Review, 65,* 182–188.

Darity, W., & Young, W. (1995). IS-LM: An inquest. *History of Political Economy, 27,* 1–41.

de Vroey, M. (2000). IS-LM à la Hicks versus IS-LM à la Modigliani. *History of Political Economy, 32,* 293–316.

de Vroey, M., & Hoover, K. D. (2004). *The IS-LM model: Its rise, fall, and strange persistence.* Durham, NC: Duke University Press.

Hansen, A. H. (1953). *A guide to Keynes.* New York: McGraw-Hill.

Hicks, J. R. (1937). Mr. Keynes and the "classics": A suggested interpretation. *Econometrica, 5,* 147–159.

Hicks, J. R. (1976). Some questions of time in economics. In A. M. Tang, F. M. Westfield, & J. S. Worley (Eds.), *Evolution, welfare, and time in economics: Essays in honor of Nicholas Georgescu-Roegen* (pp. 135–151). Toronto, Ontario: Lexington Books.

Hicks, J. R. (1980–1981). IS-LM: An explanation. *Journal of Post Keynesian Economics, 3,* 139–154.

Keynes, J. M. (1936). *The general theory of employment, interest and money.* New York: Harcourt, Brace & World.

Keynes, J. M. (1937). Letter to J. R. Hicks. In *Collected writings of John Maynard Keynes, Vol. XIV, hrsg. von Donald Moggridge: The General Theory and after. Part II: Defence and development* (pp. 79–81). Cambridge, UK: Cambridge University Press.

Kreisler, P., & Nevile, J. (2002). IS-LM in macroeconomics after Keynes. In P. Arestis, M. Desai, & S. Dow (Eds.), *Money, macroeconomics and Keynes*: *Essays in honour of Victoria Chick* (Vol. 1). London: Routledge.

Leijonhufvud, A. (1967). Keynes and the Keynesians: A suggested interpretation. *American Economic Review, 57,* 401–410.

Leijonhufvud, A. (1969). *Keynes and the classics: Two lectures on Keynes' contribution to economic theory.* London: Institute of Economic Affairs.

Minsky, H. P., & Ferri, P. (1989). The breakdown of the IS-LM synthesis: Implications for post-Keynesian economic theory. *Review of Political Economy, 1,* 123–143.

Modigliani, F. (1944). Liquidity preference and the theory of interest and money. *Econometrica, 12,* 45–88.

Robinson, J. (1962). *Economic philosophy.* Chicago: Aldine.

Robinson, J. (1971). *Economic heresies.* New York: Basic Books.

Samuelson, P. A. (1948). *Economics.* New York: McGraw-Hill.

# 34

# Economic Instability and Macroeconomic Policy

John Vahaly
*University of Louisville*

Macroeconomic instability and the business cycle are generally understood as changes in output or gross domestic product (GDP), unemployment, and inflation rates. The economy has a long-run growth path that is subject to short-term macroeconomic demand and supply shocks that push GDP away from its long-run potential or trend growth rate. Smith and the classical tradition that followed believed a hands-off approach was the correct policy stabilization to pursue when such short-term output disturbances arose. This reflected classical emphasis on long-run growth as a supply process that was best left to private entrepreneurial activity. Furthermore, private market economies would automatically self-correct through appropriate wage and price adjustments. Recessions, characterized by "gluts" of commodities and workers, would produce downward pressure on market prices and wages. Deflation continued until the economy had returned to full output.

Prior to the twentieth century, the classical school's reliance on self-correcting private markets, a commodity gold standard, and promotion of free trade broadly defined macroeconomics. Fiscal policy was viewed as the means by which government provided necessary public goods and services, with deficit spending occurring because of either war or weakness in the economy that lowered government tax receipts. Monetary policy for the United States, operating without a central bank until 1914, similarly was not considered a macroeconomic stabilization tool. Financial difficulties before and during the Great Depression were to change that.

## The Creation of the Federal Reserve System

Modern study of economic instability in the United States begins with the Great Depression. Prior to the 1930s, economists treated macroeconomic instability as a difficulty best resolved by private markets. Indeed, the U.S. financial system had even operated for most of the pre-Depression period without a central bank. There had been two experiments with a Bank of the United States but, for a variety of reasons, neither survived its initial 20-year charter. During the nineteenth century, the United States had experimented with a number of currency and banking regimes. With the gold standard officially adopted in 1900, the United States entered the new century without a central bank. The next financial panic in 1907 convinced Congress that the economy did need a central banking authority to ensure the soundness of the country's banking system. The Federal Reserve Act of 1913 created a system of 12 regional Federal Reserve Banks, with the New York bank assuming control of monetary policy. The U.S. banking system now had a lender of last resort.

## The Great Depression and the Origin of Macroeconomics

The causes and consequences of the Depression continue to be debated. A recession started in the summer of 1929 that was soon followed by the start of the collapse of the

U.S. stock market. In 1930, the United States decided to raise tariffs, soon matched by trading partners, in a mistaken attempt to protect import-competing jobs. More important, the United States experienced a series of financial panics and bank runs that would culminate with a bank holiday in 1933 as the newly elected President Roosevelt closed all banks to end withdrawal of money from the economy. Set up as a result of the financial panic of 1907, the Federal Reserve failed during its first major financial crisis.

This was understood by John Maynard Keynes in Great Britain. The Keynesian approach to dealing with the economic problems of the 1930s, both for Britain and the United States, was to rely on new deficit spending to make up for inadequate private sector spending by consumers and businesses. In the United States, policy emphasis on using fiscal policy was to continue for decades to come. This reflected a major concern with expansionary monetary policy: There might be inadequate demand for newly created bank reserves arising from Federal Reserve open-market bond purchases.

Keynesian theory, originally a depression-economy model, emphasized short-run demand stabilization through fiscal policy using either new government spending or deficit-financed tax cuts. The basic idea was to supplement private consumer sending (C) and business investment spending (I) with either direct government spending (G) or tax cuts (T in dollars or t in tax rates). Keynes explained how a fiscal stimulus, by increasing autonomous spending, would increase GDP by a multiple of the new deficit spending. The year following publication of Keynes's *General Theory,* economist John Hicks expanded Keynesian analysis to explicitly include a monetary sector and the interest rate. These additions made the Hicksian IS-LM model a standard part of macroeconomics. As with the Keynesian model, however, it was incomplete. It had no explicit means of analyzing either short-run supply or the price level. Because neither of these were problems in the 1930s, their omission is understandable. Subsequent macroeconomic events would remedy these shortcomings. Following the 1930s, however, the debate over countercyclical macroeconomic policy began in earnest.

## The Rise and Fall of the Keynesian Consensus, 1961 to 1973

The 1960 presidential election, at least in then–Vice President Richard Nixon's view, was determined by the 1960 recession. John F. Kennedy ran that year promising to enact a deficit-financed tax cut that would help restore full employment. Kennedy's victory brought the Keynesian "new economics" of activist demand management to the United States. The next recession was a decade away. Even the popular media in the 1960s wondered if the Keynesian macroeconomists who came to rise in 1961 with the new Kennedy administration had finally ended the business cycle. While economists debated its meaning and merits, Keynesian macroeconomics seemed to offer the option of fine-tuning the economy. Short-term demand management was explained as making a choice between relatively low unemployment but high inflation or relatively low inflation but high unemployment. This Phillips curve trade-off was the basis for the first ever governmental attempts at macroeconomic stabilization. Using the Keynesian spending (C + I + G) and the IS-LM models, Keynesians showed how monetary and fiscal policy could keep the economy on course. The 1960s was the longest economic expansion in U.S. history up to that time. However, as unemployment fell during the decade, inflation began rising and became a major domestic macroeconomic issue in the early 1970s.

## The Great Inflation, 1970 to 1981

From inflation rates of 1% to 2% in the early 1960s, an overheated U.S. economy was experiencing over 5% inflation by the end of the 1960s and into 1970. The U.S. economy had a relatively mild recession in 1970 with unemployment rising from a 1969 low of 3.4% to 6.1% by the end of 1970. As the economy slowed, it seemed reasonable to expect inflation to fall as excess demand was reduced. Somewhat surprisingly, inflation held steady. CPI inflation rose 5.5% in 1969 and was 5.7% in 1970 despite the weakening economy. Facing reelection in 1972, President Richard Nixon decided to impose a temporary 60-day wage and price freeze in August 1971.

The Nixon controls were an experiment with an incomes policy to deal with inflation. Largely voluntary, the Nixon program of two freezes and four phases of lessening degree of price and wage restraint was controversial. Predictably, shortages arose in a variety of markets, most notably in gasoline. However, President Nixon was reluctant to press for an alternative, which seemed to require monetary or fiscal restraint that might finally lower prices with the possibility of a worse recession during a presidential election. Mr. Nixon remembered the role a weak economy played in his 1960 election loss. A positive view of President Nixon's Phases I–IV controls was that they represented an attempt to help slow the wage–price spiral by signaling to U.S. business and workers that the federal government was going to reduce inflation by in essence making it illegal. Because the program was largely voluntary, it relied on a downward adjustment of inflationary expectations to ensure its success. People had to expect less inflation for a timely reduction in the wage–price spiral.

Unfortunately for U.S. efforts to lower inflation, other events interceded. In 1973, war in the Middle East resulted in an oil embargo engineered by the Organization of Petroleum Exporting Countries (OPEC) that had the

economic effect of quadrupling crude oil prices from $3 to $12 a barrel. Because oil is a primary source of energy, this meant demand increases for oil substitutes also pushed their prices upward, creating a major spike in energy prices. This adverse supply shock was a new challenge for U.S. policy makers. While even Mr. Nixon once famously declared "We are all Keynesians now," it was not clear to Keynesians what should be done about the effects of the adverse supply shock OPEC had created.

## Stagflation

In 1974, the U.S. economy did not seem to be operating according to the explanations of economists. The demand-side Keynesian model and its associated Phillips curve predicted two alternative economic states: excess demand (low unemployment and high inflation) or inadequate demand (high unemployment but low inflation)—the Phillips curve trade-off. But 1974 saw inflation rise from 6.2% to 11.0% while unemployment increased from 4.9% to 5.6% as the U.S. economy slipped into a recession. For 1975, the unemployment rate was 8.5%. A new term was coined by Paul Samuelson to describe this seemingly inexplicable situation—*stagflation.*

As these developments were unfolding in 1974, the previous decade's claims that the business cycle had been tamed were discarded. For policy makers, there were few options. Fiscal policy was effectively paralyzed in 1974 by the Watergate investigation. That shifted emphasis to the Federal Reserve (the Fed), then chaired by Arthur Burns. Chairman Burns was repeatedly asked by Congress and others to take action to rebalance the economy. Burns knew his options were limited to altering the money supply and credit conditions, which did not, and could not, reverse the impacts of the OPEC oil price shock.

Keynesian stabilization policies worked only on the demand side. Restrictive monetary policy could lower inflation, but the accompanying output decline would raise unemployment. The reverse was true for easy monetary policy that would push prices higher as unemployment fell. As stagflation pushed both unemployment and inflation higher, economists recognized the need to add short-term supply factors to macro models. These efforts created the aggregate demand and short-run aggregate supply model that illustrated the difficulties of the mid-1970s. The Phillips curve was also reinterpreted to include labor (supply-side) responses to inflation. But the new theory did not resolve the Fed's dilemma.

Bowing to the inevitable, the Fed first took action against inflation, and by 1975, it had dropped from 11% to 9.1% (5.8% in 1976). One economic cost of the Fed's action was the worst recession since the Depression. This contraction lasted from the fourth quarter 1973 (1973:4) to 1975:1, a 16-month recession that saw a 1975 monthly unemployment rate of 9.0%. Inflation had been reduced, but faith in fine-tuning the economy was one of the casualties.

## Stagflation: Part 2

In 1979, the United States had a new president, Jimmy Carter, but another supply shock. That year, OPEC again exercised its monopoly power and pushed crude oil prices over $30 a barrel. The impact in the United States was now predictable. In 1979, inflation averaged 11.3%, rising to 13.5% in 1980, a presidential election year. Unemployment also rose from 5.8% in 1979 to 7.1% in 1980 as the U.S. economy again slipped into a recession. Once again, the United States faced stagflation, but the numbers were bigger and uncertainty remained about what should be done about this situation.

The difficulty stagflation posed for the U.S. economy in the 1970s arose from its cause. Traditionally, inflation had been viewed as a problem of "too many dollars chasing too few goods." In the 1960s, the "too many dollars" or "too much spending" had raised demand-side inflation to over 5%. But the inflation problems in the 1970s arose from the supply side or the "too few goods" source of higher prices. It even appeared that causation was running in reverse as high inflation was causing the recession. As the energy price shock worked its way through the economy, nearly all prices were pushed up. U.S. consumers, no better off, could no longer afford what they previously purchased. GDP declined, causing unemployment to rise. By 1980, it appeared necessary to take some type of decisive action against the problem of double-digit inflation.

## The 1980 Presidential Election

With double-digit inflation as the major economic issue, incumbent Jimmy Carter and Republican nominee Ronald Reagan offered different policies to voters. Carter had appointed a new Federal Reserve chairman in 1979, Paul Volcker, in part to reestablish the Fed's credibility in dealing with excess inflation. Volcker knew this was his primary task. However, Mr. Carter did not want to put the economy through another long recession as had occurred during the 1973 to 1975 episode. Fed policy under a new Carter administration would be one of gradual monetary restraint to slow inflation.

Candidate Reagan offered something more appealing to voters. When running for his party's nomination, Mr. Reagan offered four economic platform promises: lower inflation, more jobs, a balanced federal budget, and an increase in defense spending (higher government spending, G). This was to be accomplished by a variety of initiatives, including market deregulation where possible and a major reduction in federal income tax rates. During the primaries, Reagan's rival but soon-to-be vice presidential nominee, George H. W. Bush, famously called the promise of a balanced budget with programs of increasing spending while cutting taxes "voodoo economics." The label that did stick to the Reagan program, however, was *supply-side economics.* This name came from the prediction

that lower federal income taxes would raise after-tax incomes for U.S. workers, thus encouraging additional work and more output. While the details were left unclear, Arthur Laffer produced a famous diagram that seemed to show that tax rate cuts, if they actually could boost work incentives, would eliminate the "too few goods" problem of inflation. The cure for U.S. inflation appeared painless.

Following the 1980 election, Congress and Mr. Reagan enacted the Kemp-Roth tax cut and began the defense buildup. The Federal Reserve, however, began ringing alarm bells. Volcker and many others viewed supply-side fiscal policy, whatever its ideological merits, as expansionary. Tax rate cuts coupled with significant government spending increases would increase the "too much spending" problem of inflation. Volcker concluded that expansionary fiscal policy would have to be offset by very restrictive monetary policy. In 1981, the Fed acted. Aggregate demand was reduced, as restrictive monetary policy more than offset the federal fiscal expansion. By 1982, inflation had dropped from 10.3% to 6.2% (3.2% in 1983). The cost of this success was another recession, this time even worse than the 1973 to 1975 slowdown, with unemployment reaching a monthly high of 10.8% in 1982. The 1981:3 to 1982:4 recession was the worst macroeconomic performance since the 1930s. A new term, *disinflation,* was coined to describe the successful lowering of inflation. Inflation had been lowered to 3% to 4% but would fall no farther in coming decades in part to avoid a repeat of the damage done by the 1981 to 1982 monetary restraint.

### The Great Bull Market and the Twin Deficits

As the U.S. economy started recovering from the 1981 to 1982 "Volcker Recession," the stock market also started to move up. The Dow Jones Industrials reached a low of 777 in August 1982. The subsequent recovery of the economy began a rise in equity values that was twice interrupted in the next 25 years. The first, "Black Monday" (October 19, 1987), occurred just as the Fed chairmanship changed from Paul Volcker to Alan Greenspan. The second and longer retrenchment was associated with the tech stock or dot-com bubble in 2000. From the 1982 low, the Dow rose to 11,722 by 2000, an increase of more than 1,500% over this 18-year period. The market subsequently dropped to 7,286 in October 2002. As the economy recovered from the 2001 recession, both the equity and housing markets moved toward historic highs. The Dow Jones reached this most recent high of 14,146 in October 2007. In the following year, the Dow Jones Industrial Average (DJIA) lost nearly half its value as the U.S. economy began a dramatic slowdown as the financial system entered a crisis period brought on by the housing market collapse.

In 1980, the U.S. national debt was $0.9 trillion, and the United States was the world's largest creditor nation. By the end of the decade, new deficit spending had raised the national debt to $3.2 trillion, and the United States had become the world's largest debtor country resulting from a decade of trade deficits. These two deficits are linked to saving rates in both the United States and abroad. Both debts would continue to expand after the 1980s.

## The 1990 to 1991 Recession and the Start of the U.S. Housing Boom

The U.S. economy did not fully recover from the 1981 to 1982 recession until the late 1980s. The monthly unemployment rate did not fall below 6% until September 1987 and did not reach 5% until March 1989. In 1990, however, the U.S. economy began to slow. One factor was a decline in the housing market following the collapse of the Savings and Loan banking sector. In the summer of 1990, unemployment started to rise, and by the end of the year, it stood at 6.3%. Officially, the recession started in July (1990:3) and lasted until March 2001, spanning 8 months. There would not be another recession for 120 months, the longest cyclical expansion in U.S. history.

The short-term policy response to the 1990 to 1991 recession was a gradual series of interest rate decreases under the guidance of Fed Chairman Greenspan, who had replaced Paul Volcker in 1987. The Federal Funds ("fed funds") interbank loan rate stood at 8% when the recession started in 1990. By the summer of 1991, it was 6%, falling to just 3% by the end of 1992. The Fed pursued gradual monetary ease to stimulate spending and help the economy recover.

This expansion also marked an increase in investment spending on the U.S. housing stock. What began as a typical housing recovery would later culminate in a speculative rise in U.S. housing prices. From 1991 through 1996, average U.S. housing prices rose at an annual rate between 2% and 4%, roughly matching inflation. From 1998 through 2006, the rate of increase doubled. At the end of the boom, housing price rises were 9% in 2004 and 2005. Once the bubble had burst, home prices rose a modest 3.3% in 2006 and fell –1.3% in 2007. The bursting of this asset price bubble was to have major macroeconomic consequences that were largely unforeseen.

## The Dot-Com Stock Market Bubble and the 2001 Recession

The DJIA reached its 2000 peak of 11,722 on January 14, 2000, falling to a low of 7,286 by October 9, 2002. This loss of financial wealth contributed to weakness in consumer spending. More important was the collapse of business investment spending, especially for computers and software, during 2001.

The 8-month 2001 recession ran from March until November of that year. In January 2001, the unemployment rate was a low 4.2%. Following the end of the recession, it peaked at 6.3% in June 2003. In response to this relatively slow recovery, the Fed once again adopted a policy of monetary

ease. 2001 started with the fed funds rate at 6%. It was progressively lowered to 1% by the summer of 2003 as the Greenspan Fed tried to move the economy back to full employment. By the summer of 2005, the unemployment rate had reached 5% as the Fed ended its policy of easy monetary policy. As 2005 began, the fed funds rate was 2.25% and would reach 5.25% in late 2006.

## The Collapse of the Housing Bubble

The National Bureau of Economic Research (NBER), established in 1920, dated a recession as starting in December 2007. The recognition of this downturn came in December 2008. This somewhat unusual delay was caused by the mixed signals being generated by macroeconomic aggregates. While real GDP grew for the first two quarters of 2008, total employment had peaked in December 2007 and fell thereafter. The employment data, and other economic indicators, convinced the NBER's Business Cycle Dating Committee that the actual recession had started December 2007.

As noted, the U.S. housing market began expanding following the 1990 to 1991 recession. The housing boom continued even during the 2001 recession with residential investment declining only in 2001:4, after the trough of the recession was reached. Home prices kept rising and even accelerated through 2005. Such asset price trends cannot be maintained, and the inevitable slowdown began in 2006. Home prices began falling as that sector was now characterized by both excess supply and falling demand. This had a variety of negative impacts on the economy, including a major decrease in financial wealth as the stock market plunged. The Dow Jones peaked at 14,164 on October 9, 2007. During 2008, the market lost one third of its value by this measure. Not only was financial wealth cut as equities dropped, but many now saw the value of their home falling as well. With this twin wealth shock, consumer confidence in the economy dropped to historic lows.

The business and financial sectors were also adversely affected. The construction industry was among the first casualties as new residential construction stalled. Worse, the financial system seemed unable to cope with the magnitude of the financial distress the collapsing housing bubble had produced.

The list of factors associated with the rise and fall of the U.S. housing market is remarkably long. Alan Greenspan acknowledged he had placed too much confidence in the self-regulatory character of complex financial markets and financial derivatives. While the boom unfolded, money was drawn to the United States from around the world. To finance the levitating housing market, first investment banks then government-sponsored enterprises (GSEs) Fannie Mae and Freddie Mac developed new sources of funding. First, this process relied on the selling of bundles of mortgages, or securitized mortgage debt, to investors eager to benefit from the housing boom. Given the demand for these mortgage-backed securities (MBS), it was not surprising loan standards were loosened as "subprime" or risky lenders were given mortgages that were quickly resold into an ever-growing mountain of paper wealth.

This expansion of housing credit occurred against a backdrop of financial deregulation, questionable rating of the risks associated with MBS and related financial contracts, complex financial investments that were usually marketed through in a lightly regulated "shadow banking system," charges of predatory lending and borrowing, and an apparent belief that the housing price rise was too good an opportunity to miss. There was also a government commitment to promote home ownership. Thus both Fannie Mae and Freddie Mac were encouraged to help provide the financing necessary to help people buy a home while the boom was in progress. Perhaps one positive, if transitory, benefit of the bubble was a 69.2% home ownership rate in 2004, the highest in U.S. history.

As the housing collapse started to have macroeconomic effects in late 2007, the economic impacts were historic. In 2008, the economy experienced the largest corporate bankruptcy in history (insurance company AIG), the largest bank failure in U.S. history (Lehman Brothers), and the largest Savings and Loan failure in U.S. history (Washington Mutual). All major investment banks failed, merged, or converted themselves into traditional deposit banks. Major U.S. commercial banks asked for Treasury funding support through new fiscal programs.

## Monetary Policy and the 2007 Recession

The macroeconomic policy response to this recession was unprecedented in its size and scope. Monetary policy in 2008, under Chairman Bernanke, was the most expansionary since World War II (WWII). The Fed's emphasis remained on the stability of the banking sector and to a lesser extent on inflationary concerns as total spending in the economy declined. As with Japan during the 1990s, concern was expressed over the appearance of deflation further reducing Fed worries over inflation.

Traditionally, monetary policy is first seen as the Federal Reserve Board, acting through the Federal Open Market Committee (FOMC), altering credit conditions in the economy using open-market operations. Predictably, following 9/11 and the 2001 recession, the Fed lowered the fed funds rate to 1% by mid-2003. Alan Greenspan, Fed chairman, made clear his policy goal was overall stimulus of the economy. However, 30-year fixed rate mortgage rates, which had been 8% in 2000, fell below 6% during the same period. Therefore, further stimulus was given to the U.S. housing market, whose growth appeared to be feeding on itself. By mid-2006, monetary restraint had pushed the fed funds rate back to 5.25%. Against this background, Greenspan was criticized for following traditional easy monetary policy in a weak economy that nonetheless was experiencing a residential and commercial real estate boom.

When the housing price bubble began to lose air in 2005 and 2006, the Fed began to take new and dramatic action. In December 2007, new Fed Chairman Ben Bernanke created the Term Auction Facility (TAF), the first of seven completely new lending operations that would provide liquidity (money) to a variety of financial institutions and markets that had stopped working efficiently. Lenders had decided to significantly reduce traditional lending as the "credit crunch" spread. There was no longer any market for MBS or other high-risk alternatives. The Fed even allowed short-term borrowing by financial institutions that were able to use some higher grade MBS as collateral and finally agreed to purchase some MBS itself. All through 2008, the Fed kept creating new auctions and programs of providing credit to the economy. The Fed also pursued traditional easy monetary programs. At the end of 2007, the Fed funds rate was 4.25%. By the end of 2008, it was 0% (officially a range of 0% to 0.25%).

The immediate question this zero-interest rate policy posed was, had the United States run out of monetary policy options just as the NBER announced a recession had started in December 2007? Certainly fiscal policy could be used to stimulate spending, but had monetary policy run aground?

While the Fed was lowering the nominal fed funds rate to zero, it simultaneously pursued a program of "quantitative" monetary ease. The Fed accomplished this in two ways. The first was continuation of traditional monetary expansion. The monetary base doubled in 2008, increasing from $855 billion in December 2007 to over $1,728 billion by December 2008. This expansion was not needed to lower the fed funds rate to zero; rather, its aim was to help recapitalize the U.S. banking system. Inflationary concerns were put on hold as the Fed kept adding to bank reserves. For the banking sector, this dramatically increased the bank holdings of excess reserves as the Fed kept adding liquidity to the economy.

The second policy still available to the Fed was to alter its holdings of U.S. Treasury securities. The Fed could provide additional monetary stimulus to borrowers by also lowering long-term interest rates. Buying long maturity Treasury bonds pushed up their prices, which pulled their yields or rates down, putting downward pressure on related long-term rates, mortgage rates in particular.

The issue of policy credibility was also important for the Fed's commitment to recovery. All through 2008, the Fed signaled that low interest rates and market support programs would be used as long as required by the economy. With support from fiscal policy actions, the Fed had embarked on the most expansionary policy in its history.

## Fiscal Policy and the 2007 Recession

Fiscal ease was viewed as an inevitable policy response to the same problems that preoccupied the Fed. Fiscal policy was used to help both the overall economy and financial markets and institutions. In early 2008, the Economic Stimulus Act of 2008 sent $150 billion into the private economy as temporary tax rebates. The worsening economy and the fact the tax change was temporary meant that much of this additional income was not spent by consumers. By the fall of 2008, fiscal action moved into new country as the Treasury Department asked for a total of $700 billion to assist the economy.

Fiscal policy embarked on programs that would produce $1 trillion deficits. Both stabilization policies tried to slow the economy's retrenchment following the collapse of the housing market. As in the 1930s, the major problem facing the U.S. economy was the potential collapse of its banking and financial sectors. The evolution of financial intermediation in the United States was thought to have been one of increasing sophistication that reduced risk while generating income. As noted previously, this view, widely expressed prior to the collapse of the housing market, was incorrect. Rather than reducing risk, debt securitization, related financial derivatives, and inadequate market supervision magnified the temptation to assume ever-growing risk. After all, the economy had grown for 25 years with no major economic difficulties. As with previous asset bubbles, the U.S. housing market boom did finally come to an end. The macroeconomic costs of repairing the damage done by this largest ever bubble will take years to complete.

## The U.S. National Debt and Foreign Trade Debt

While both monetary and fiscal policy were being used to offset spending declines arising from the 2007 recession, the twin deficits that emerged in the 1980s remain unresolved. Deficit spending, except for 3 years at the end of the 1990s, has continued. Given the combination of a weak economy and expansionary fiscal policy, deficit growth will accelerate. Thus, the short-term projection for the national debt is significant growth until full employment is restored.

The cause of the persistent trade deficits over the past quarter century arise from another domestic macroeconomic imbalance, here between saving and investment. U.S. national saving is so low that the United States must finance domestic investment spending with the help of foreign savings. Given expanding federal budget deficits, there is additional pressure on the trade deficit despite reductions in private investment spending for buildings and equipment. Most observers therefore conclude that U.S. trade deficits, and their financial liabilities, will continue until the U.S. saving rate can be increased. This adjustment can be done in an orderly fashion by promoting both private and public sector saving. It could also happen as a matter of economic necessity should the United States no longer be in a position to finance substantial federal

debt and private investment spending. That undesirable outcome is being considered if only to promote thinking about more desirable resolutions. Debt does not have to be zero, but it must be manageable.

## The Future of Macroeconomic Policy

The debate over stabilization policy displays its own cyclical features. The classical tradition of laissez-faire was displaced by the Great Depression and the demand management policies of *The General Theory*. Such countercyclical policies were applied in the 1960s when it almost seemed that the business cycle had been managed away. Theory lapses and supply shocks demonstrated the weaknesses of the Keynesian consensus in the 1970s. At best, "coarse tuning" was the new standard. A variety of new conservative macroeconomic schools emerged that encouraged reconsideration of the wisdom of laissez-faire policies for the business cycle. Despite academic debates about such issues, stabilization efforts continued after the 1970s. Volcker's Fed acted to slow inflation in the early 1980s, and Alan Greenspan's first challenge, the October 1987 stock market crash, saw the Fed act to prevent additional damage from this event. Greenspan continued to purse countercyclical policies in the two recessions that occurred during his tenure. Greenspan was even criticized for pursuing overly expansionary policies in 2003 to 2004 as the housing market boom continued. However, Greenspan said the error he made was in misjudging the stability and risk-management capabilities of financial markets. Furthermore, it was not clear that it was the Fed's responsibility to prevent the occurrence of asset price bubbles.

If anything, Bernanke's Fed has been the most active ever in dealing with the difficulties the U.S. economy was facing. Along with the recession, the Fed had to operate in an economy that was hit by a major equity-price shock as stock markets around the world dropped dramatically. For 2008, U.S. stock prices, as measured by the level of the Dow Jones Industrials, fell by one third. Across all equity markets, trillions of dollars of financial wealth were erased. This negative wealth effect affected private-sector consumer and business spending. Further eroding consumer confidence was the apparent fear of the federal government as it pushed for new emergency deficit spending to keep collapsing financial markets operating. The fiscal policy response was initially one of additional spending and tax cuts, and in relatively large amounts, before taking on regulatory reform. The combination of remarkably expansionary fiscal and monetary policy with structural reform of financial markets is being counted on to repair the damage from the greatest financial bubble in history. Macroeconomic events of the 1930s were somewhat similar, with financial turmoil, falling equity prices, and rising unemployment. It took the expansionary macro policies of WWII to finally pull the economy back to full employment. In 2009, such spending policies are in place because of the lessons of that era.

## References and Further Readings


Abrams, B. A. (2006). How Richard Nixon pressured Arthur Burns: Evidence from the Nixon tapes. *Journal of Economic Perspectives, 20*(4), 177–188.

Ball, L., & Mankiw, N. G. (2002). The NAIRU in theory and practice. *Journal of Economic Perspectives, 16*(4), 115–136.

Chari, V. V., & Kehoe, P. J. (2006). Modern macroeconomics in practice: How theory is shaping policy. *Journal of Economic Perspectives, 20*(4), 3–28.

Crowe, C., & Meade, E. E. (2007). The evolution of central bank governance around the world. *Journal of Economic Perspectives, 21*(4), 69–90.

Elsby, M. W. L., Michaels, R., & Solon, G. (2009). The ins and outs of cyclical unemployment. *American Economic Journal: Macroeconomics, 1*(1), 84–110.

Feldstein, M. (2008). Resolving the global imbalance: The dollar and the U.S. saving rate. *Journal of Economic Perspectives, 22*(3), 113–125.

Friedman, M. (2005). A natural experiment in monetary policy covering three episodes of growth and decline in the economy and the stock market. *Journal of Economic Perspectives, 19*(4), 145–150.

Galí, J., & Gertler, M. (2007). Macroeconomic modeling for monetary policy evaluation. *Journal of Economic Perspectives, 21*(4), 25–46.

Goodfriend, M. (2007). How the world achieved consensus on monetary policy. *Journal of Economic Perspectives, 21*(4), 47–68.

Gordon, R. J. (2008). *Macroeconomics* (11th ed.). Boston: Pearson, Addison Wesley.

Hoshi, T., & Kashyap, A. K. (2004). Japan's financial crisis and economic stagnation. *Journal of Economic Perspectives, 18*(1), 3–26.

Jorgenson, D. W., Ho, M. S., & Stiroh, K. J. (2008). A retrospective look at the U.S. productivity growth resurgence. *Journal of Economic Perspectives, 22*(1), 3–24.

Landefeld, J. S., Seskin, E. P., & Fraumeni, B. M. (2008). Taking the pulse of the economy: Measuring GDP. *Journal of Economic Perspectives, 22*(2), 193–216.

Mehrling, P. (2002). Retrospectives: Economists and the Fed: Beginnings. *Journal of Economic Perspectives, 16*(4), 207–218.

Romer, C. D., & Romer, D. H. (2004). Choosing the Federal Reserve chair: Lessons from history. *Journal of Economic Perspectives, 18*(1), 129–162.

Snowdon, B., & Vane, H. R. (2005). *Modern macroeconomics.* Cheltenham, UK: Edward Elgar.

Svensson, L. E. O. (2003). Escaping from a liquidity trap and deflation: The foolproof way and others. *Journal of Economic Perspectives, 17*(4), 145–166.

van Ark, B., O'Mahoney, M., & Timmer, M. P. (2008). The productivity gap between Europe and the United States: Trends and causes. *Journal of Economic Perspectives, 22*(1), 25–44.

Woodford, M. (2007). The case for forecast targeting as a monetary policy strategy. *Journal of Economic Perspectives, 21*(4), 3–24.

Woodford, M. (2009). Convergence in macroeconomics: Elements of the new synthesis. *American Economic Journal: Macroeconomics, 1*(1), 267–279.

# 35

# Fiscal Policy

Rosemary Thomas Cunningham
*Agnes Scott College*

*Fiscal policy* refers to the government's use of spending and tax policies to influence the economy. When the government increases its spending for defense purposes or raises personal income tax rates, it affects the total level of spending in the economy and, hence, will affect the overall macroeconomic activity of a nation measured by such factors as gross domestic product (GDP), employment, and inflation. This is true for almost any change in spending or taxes. Any change in government spending or taxes will also affect the government's budget deficit. An increase (decrease) in spending or a decrease (increase) in taxes will increase (decrease) the government's budget deficit for a given state of the rest of the economy. Because the government must borrow by selling U.S. Treasury securities to finance its deficits, any increase or decrease in the government's deficit will affect the market for loanable funds and interest rates, which then feeds back on GDP, employment, and inflation.

This chapter looks at the impact of government's spending and tax policies, discussing the ways in which it can affect, and has affected, the economy. It also presents various arguments for greater reliance on fiscal policy to increase employment, as well as discusses the problems that increased government deficits may impose on future generations.

## Federal Government Spending

Economists categorize government spending in two types. First, there is government purchases of goods and services. Government purchases of food, military goods, and other goods needed for consumption purchases, as well as purchases of investment goods, such as buildings and the building of bridges, are included. On the other hand, the government also spends on social insurance programs, such as Food Stamps and Medicare, which are primarily transfers of income from taxpayers to needy citizens. These are referred to as *transfer payments.*

Federal, state, and local governments all purchase goods and services and have the potential to affect economic activity. According to the 2008 *Economic Report of the President,* over the past 46 years, total government spending on goods and services, as well as federal spending by itself, has fallen as a percentage of GDP while state and local spending has increased as a percentage of GDP. Total government spending was approximately 19.1% of GDP in 2007: Federal spending was 7.1% of GDP while state and local spending was 12.2%. This compares with total government spending of 21.2% of GDP in 1960, when federal spending was 12.2% of GDP and state and local spending was 9.0%.

According to Alan Auerbach (2007), since 1962 there have been substantial changes in the components of government spending. He highlights the decrease in defense spending over this time with the notable exceptions of the first half of the Reagan administration and in the post-9/11 era. However, as Auerbach explains, "entitlement spending has more than doubled as a share of GDP since the early 1960s, absorbing the 'peace dividends' provided by the conclusions of the Vietnam and Cold Wars" (p. 215).

The Congressional Budget Office (2009) provides more insight into total government spending by dividing federal government spending into discretionary spending, mandatory spending net of any offsetting receipts, and net interest. Figure 35.1 below shows how federal spending as

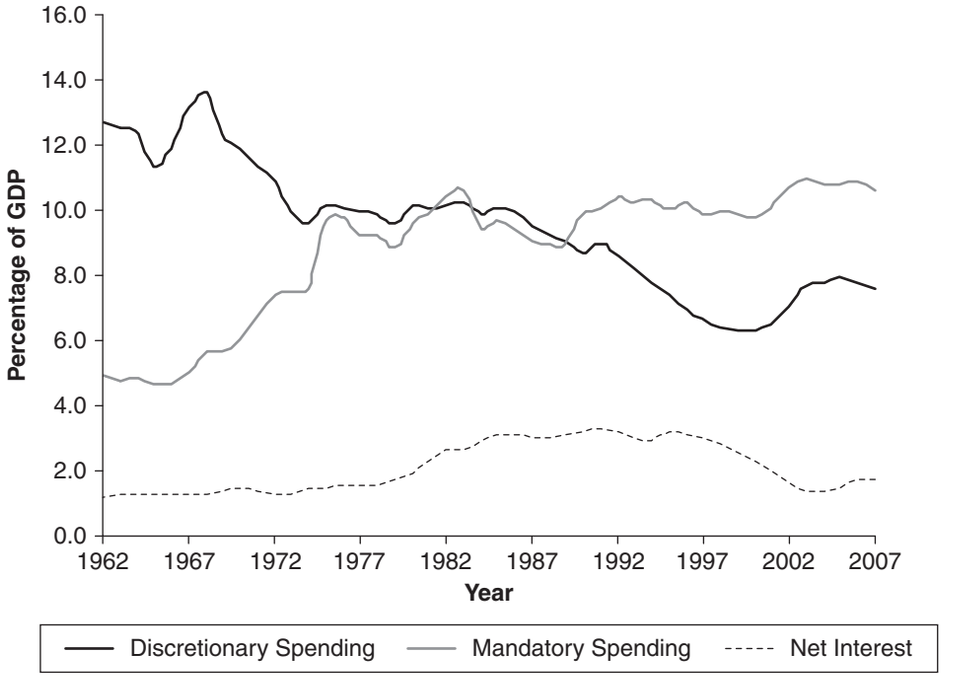


**Figure 35.1** Components of Federal Spending

SOURCE: Congressional Budget Office (2009).

a percentage of GDP has changed over time from 1962 to 2007. Discretionary spending has fallen over time, mandatory spending has risen, and net interest has varied.

*Discretionary spending* consists of items on which Congress chooses to spend, such as defense and non-defense spending, while *mandatory spending* consists of programs where spending levels are already determined as a matter of law, such as spending on Social Security, Medicare and Medicaid, and other income, retirement, and disability programs, including unemployment compensation, Supplemental Security Income, and Food Stamps, among others. Figure 35.2 shows the changing

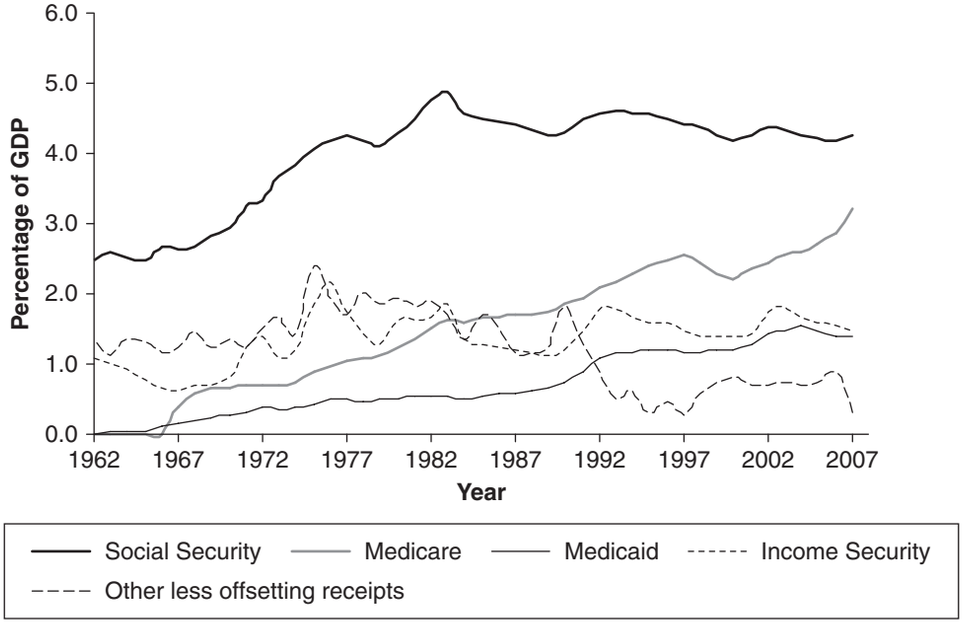


**Figure 35.2** Components of Mandatory Spending

SOURCE: Congressional Budget Office (2009).

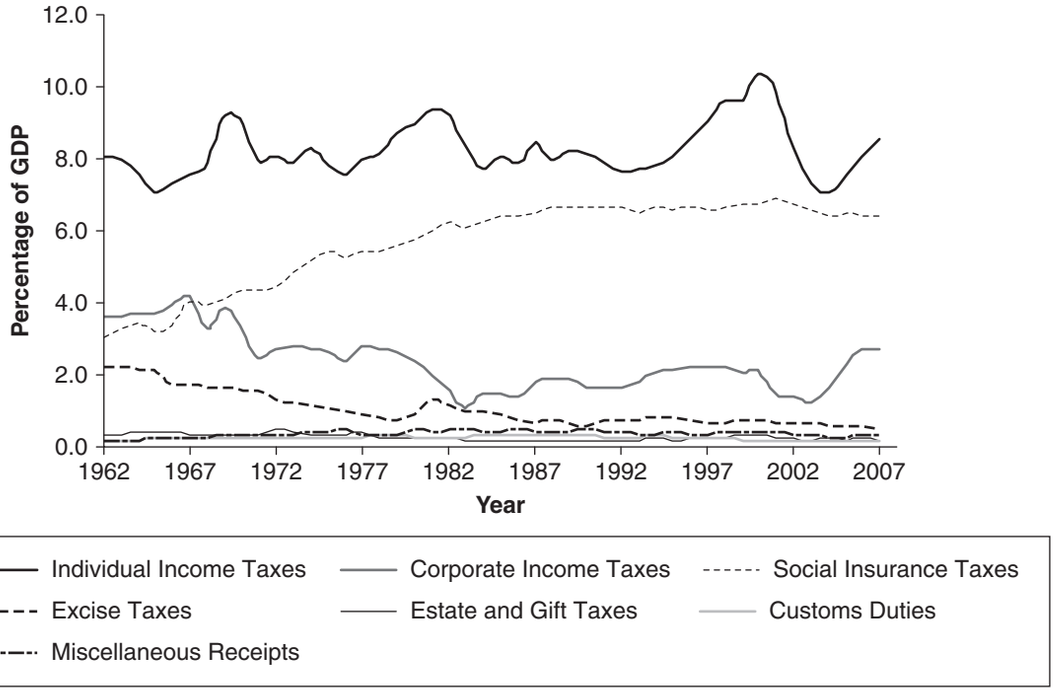


**Figure 35.3** Components of Tax Revenue

SOURCE: Congressional Budget Office (2009).

composition of mandatory spending over time. The major changes in the percentage of the components of mandatory spending as a percentage of GDP occur in both Social Security and Medicare. As the baby boomers age, these components promise to become even larger components of mandatory spending.

## Federal Tax Revenue

The federal government has several tax revenue streams, including individual income taxes, corporate income taxes, social insurance taxes (to fund programs intended to protect households from economic hardship), excise taxes (taxes on specific goods such as cigarettes), estate and gift taxes, custom duties, and other income streams. Figure 35.3 shows that individual income taxes are the most important source of tax revenue, followed by the recently more important social insurance taxes. The third most important source is excise taxes, with corporate income taxes, estate and gift taxes, and custom duties representing a small percentage of GDP. From another perspective, Figure 35.4 shows the sources of federal tax revenue in 2007. This figure again illustrates that

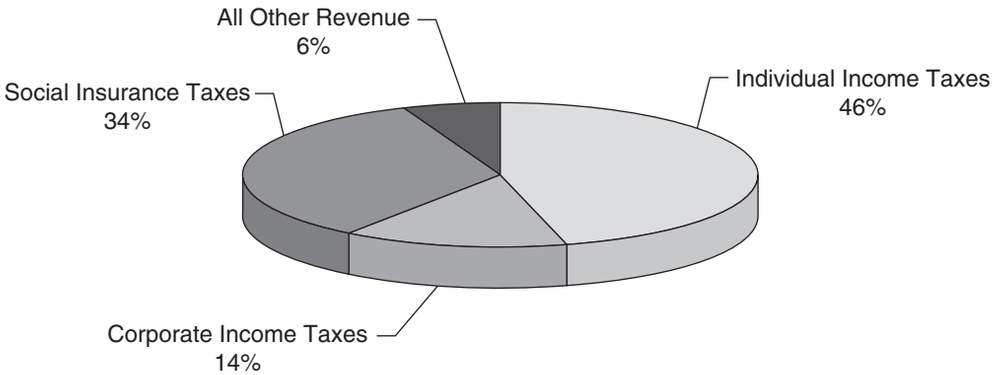


**Figure 35.4** Components of Federal Tax Revenue

SOURCE: Congressional Budget Office (2009).

individual income taxes plus social insurance taxes represent almost 80% of all federal tax revenue.

Personal income taxes in the United States are progressive, meaning that individuals with higher incomes pay a higher percentage of their income in taxes. In 2008, there were six different tax brackets with individuals paying a higher percentage of their income in taxes as their income rose. Although the income brackets change depending on whether the taxpayer is single, head of household, or married, the rates were 10%, 15%, 25%, 28%, 33%, and 35%. A report by the Congressional Budget Office to the Chair of the Finance Committee of the U.S. Senate in 2008 found that the tax system is indeed progressive in practice. The Congressional Budget Office reported that with regard to all federal taxes (not just income taxes), those individuals in the first income quintile (those whose income are among the lowest 20% in the country) pay an effective tax rate of 4.3% of income, those in the second quintile pay 14.2%, in the third pay 14.2%, and in the fourth pay 17.4%, while those whose incomes are among the 81% to 90% pay 20.3%, 91% to 95% pay 22.4%, 96% to 99% pay 25.7%, and those in the top 1% pay 31.5%.

Although it is difficult to compare tax burdens internationally, *Forbes* (Anderson, 2006) attempted to measure a Tax Misery & Reform Index that presents an international comparison of top marginal rates of taxation. The Misery scores are the sum of personal income taxes, corporate income taxes, wealth taxes, employer social security taxes, employee social security taxes, and VAT/sales taxes. From this analysis, it does not appear that Americans are particularly heavily tax burdened. Allowing for some differences within countries due to property and sales taxes, China and the original European Union-15 nations have the highest levels of tax misery, with index values ranging from 121.0 to 166.8, while the United States has an index value between 115.7 in New York and 94.6 in Texas.

## Automatic Stabilizers

Fiscal policy also provides some automatic stabilization for the economy during the business cycle. For example, government spending on unemployment compensation, Supplemental Security Income, and Food Stamps will decrease (increase) during an expansion (recession), while personal income taxes will decrease (increase) during a recession (expansion). The decrease in taxes is also more than a proportional decrease. Because the personal income tax system is progressive (i.e., individuals pay a higher proportion of their income as the income rises), it may be that the decrease in income lowers not only the total amount of taxes paid but also the proportion of income paid in taxes. So as personal income falls (rises) during a recession (expansion), fiscal policy acts as an automatic stabilizer for income by increasing (decreasing) social insurance spending and lowering (raising) taxes, thereby helping to maintain a household's income when the economy is in recession (expansion).

## Social Security and Medicare

There is a great concern that government spending on Social Security and Medicare may soon overwhelm the budget and lead to larger and larger government deficits. According to Rudolph Penner and C. Eugene Steuerle (2007), "Discretionary programs now constitute less than 40 percent of spending, whereas they were almost 70 percent of spending in 1982" (p. 1). They also explain that although today about half of federal spending goes to benefit people aged 65 and over through the Social Security, Medicare, and Medicaid programs, they warn that by the year 2030, these "programs are expected to absorb between 5 and 9 percent more of gross domestic product than they did in 2006." This is due to the increase in life expectancy and the improvement in health care. Adding to the problem is the fact that birthrates fell rapidly in the 1960s and remain low. As Penner and Steuerle explain, today's labor force, and with it the number of taxpayers, is now growing slowly, but that growth is expected to decelerate in the future.

## State and Local Government Spending and Tax Revenue

Although state and local governments may have individual income taxes, the primary ways in which they raise tax revenue is through property and sales taxes. According to the U.S. Census Bureau, in 2005 and 2006, almost 65% of all state and local tax revenue in the United States was in the form of property and sales taxes. Education, the largest single source of expenditure for state and local governments, is approximately 30% of total expenditure. It is followed by programs for social insurance and income maintenance at 22% of total expenditure and utility expenditure at 13.5%.

## Theory

Macroeconomics is a relatively new branch of economics. Although classical economists had a framework for looking at the macroeconomy, they believed that any unemployment problem would be quickly resolved and that the economy would usually function at full-employment. They argued that production was a part of the consumption process. In other words, we produce goods and services because we want to sell them for money to buy what we want to consume. If we produced too many of some types of goods and services, the prices of those goods and services would fall, and we would no longer be able to buy the goods we wanted to consume. As a consequence, we would change what we were producing, and employment would rise in the production of the higher value goods. In the view of the classical economists, the economy would tend toward full employment, and there could not be any sustained unemployment.

It was with the Great Depression and the publication of John Maynard Keynes's *General Theory of Employment, Interest and Money* that economists began to understand that there may be a role for government to play in influencing the level of output, employment, and prices. Keynes asserted that rather than supply creating demand, it was demand that creates supply, and therefore there could be a persistent oversupply of goods. There was no reason that the economy would tend toward full employment, and unemployment could be severe and persistent. Keynes argued that this was what was happening during the Depression. He also argued that the government could alleviate the unemployment by increasing demand through either increases in government spending or decreases in taxes.

## Equilibrium Output and Equilibrium Income

National output (GDP) is determined by identifying the spending in the economy by the major players: households, firms, and government. Goods bought by households are defined as consumption (C), goods bought by firms as private investment (I), and goods bought by the government as government spending (G). If the economy does not engage in international trade, total demand equals consumption plus private investment plus government spending (i.e., C + I + G). The economy will be in equilibrium when what is produced is exactly equal to what is demanded, or

$$Y = C + I + G,$$

where Y is total output produced. In this case, equilibrium output also represents equilibrium expenditure in the economy. Further, if mechanically expenditure must equal income, then Y also represents equilibrium national income.

If the economy is allowed to trade, some of the consumption, private investment, and government spending may be on goods that are not produced in the country; they may be on imports (IM). If this is the case, then these imports do not represent demand for goods produced in this country. However, it may be that foreigners will demand some goods that will be produced in this country; these are our exports (X). In an open economy, equilibrium will be achieved when domestic production equals the demand for domestic goods and services, or

$$Y = C + I + G + (X - IM).$$

There is no reason why domestic production should equal the domestic demand for goods and services at full employment. Full employment exists when workers who are willing and able to work at the going wage rate can find employment. It may be that the domestic production equals the demand for goods and services at a level of output that is less than full employment; if this is the case, the economy is in recession. Changes in government spending and taxes can help bring the economy to full employment. If an economy is in recession, the government can either increase government spending or decrease taxes (if personal taxes decrease, consumption will increase), increasing domestic demand for goods and services. Production and employment will increase. If there is inflation in the economy, the government can either reduce its spending or raise taxes, reducing domestic demand for goods and services, thereby reducing inflationary pressure. When the government uses changes in government spending or taxes to affect economic activity, it is said that the government is using discretionary fiscal policy.

## The Multiplier

According to Keynesian theory, the effect of a change in government spending or a change in consumption spending due to the change in taxes on equilibrium income will be larger than the initial change in spending. There will be a multiplied effect on output due to the initial change in spending because consumption spending itself is a function of disposable income (i.e., national income less taxes). The Keynesian view of consumption is that it is a function of disposable income. As disposable income increases, there will be a propensity to spend part of that increase and to save part of it. Given an additional dollar in disposable income, households will be likely to spend a certain percentage and to save a certain percentage. The percentage that it is likely to be spent is known as the *marginal propensity to consume* (mpc), while the percentage that it is likely to be saved is known as the *marginal propensity to save* (mps). The mpc plus the mps must equal 1.

### Government Spending Multiplier

An example can illustrate how a change in government spending can have a multiplied effect on national income. Table 35.1 shows the changes in spending and income in this example. Suppose an economy were in equilibrium at a level of output less than full employment and the government increased its spending by $100 billion. Assuming that taxes are not a function of income but are a fixed amount per person, any increase in income will also increase disposable income by that amount. However, if the marginal propensity to consume in this economy is 80%, then the initial increase in income of $100 billion will also induce an increase in consumption spending of $80 billion. This induced increase in consumption spending also raises disposable income by $80 billion, leading to a further increase in consumption spending of $64 billion, and so on. The initial increase in spending (whether government spending or consumption spending due to a change in taxes) induces various rounds of consumption spending with each round spending (and income increasing) less and less. As Table 35.1 illustrates, if the mpc equals 80% and there is an initial increase in spending of $100 billion, this will induce a total increase in consumption spending of $400 billion and will ultimately increase equilibrium income by $500 billion. In this example, the

**Table 35.1** The Government Spending Multiplier

| *Round* | *Initial Change in Spending* | *Change in Income* | *Induced Change in Consumption* |
|---|---|---|---|
| 1 | $100 | $100 | $80 |
| 2 | | $80 | $64 |
| 3 | | $64 | $51 |
| 4 | | $51 | $41 |
| 5 | | $41 | $33 |
| … | | | |
| 20 | | $1 | $1 |
| … | | | |
| All Rounds | | $500 | $400 |

multiplier is 5; with a mpc of 80%, any initial change in spending will increase equilibrium income by 5 times the initial change in spending. The value of the multiplier is

$$\text{Multiplier} = 1/(1 - \text{mpc}).$$

*The Tax Multiplier*

The government could also have a positive effect on equilibrium national income by lowering taxes by $100 billion, but the effect on the increase in national income will be smaller. The value of the tax multiplier is smaller than the value of the spending multiplier. Table 35.2 illustrates this example. If the government lowers taxes by $100 billion, disposable income will rise by $100 billion. Given the mpc of 80%, this will result in an initial increase consumption spending of $80 billion and an increase in saving of $20 billion. In contrast to the preceding example, when government spending increased by $100 billion, it is the $80 billion increase in consumption spending that starts the multiplier effect. Ultimately, the $100 billion tax cut raises equilibrium national income by $400 billion. The tax multiplier is only 4; any decrease in taxes will increase equilibrium national income by 4 times the decrease in taxes. The value of the tax multiplier is

$$\text{Multiplier} = -\text{mpc}/(1 - \text{mpc}).$$

It is negative because a decrease in taxes will increase equilibrium national income and vice versa.

*Proportional Taxes*

Relaxing the assumption that taxes are a fixed amount per person and introducing a proportional tax system (i.e., taxes are a function of income), the size of the multiplier decreases, but a change in government spending will still have a multiplied effect on equilibrium output. Continuing with the preceding example, one can determine the change in equilibrium income if the government increases its spending by $100 billion when the mpc is 80% and taxes are 10% of income. Because taxes are a function of income, disposable income equals national income minus taxes. In this case, as the government increases its spending by $100 billion, income rises by $100 billion, but now households will have to pay 10% of the increase in income in taxes leaving an increase in disposable income of only $90 billion. With an mpc of 80%, this will induce an increase in consumption spending of $72 billion. Table 35.3 shows the rounds of spending and income when a 10% proportional tax is imposed. The total change in equilibrium income is $357 billion (or approximately $36 billion in taxes and

**Table 35.2** The Tax Multiplier

| *Round* | *Initial Change in Taxes* | *Initial Change in Disposable Income* | *Initial Change in Consumption* | *Change in Income* | *Induced Change in Consumption* |
|---|---|---|---|---|---|
| 1 | –$100.00 | $100.00 | $80.00 | $80 | $64 |
| 2 | | | | $64 | $51 |
| 3 | | | | $51 | $41 |
| 4 | | | | $41 | $33 |
| 5 | | | | $33 | $26 |
| … | | | | | |
| 20 | | | | $1 | $1 |
| … | | | | | |
| All Rounds | | | | $400 | $320 |

**Table 35.3** The Proportional Tax Multiplier

| *Round* | *Initial change in Spending* | *Change in Income* | *Change in Taxes* | *Change in Disposable Income* | *Induced Change in Consumption* |
|---|---|---|---|---|---|
| 1 | $100.00 | $100.00 | $10.00 | $90.00 | $72.00 |
| 2 | | $72.00 | $7.20 | $64.80 | $51.84 |
| 3 | | $51.84 | $5.18 | $46.66 | $37.32 |
| 4 | | $37.32 | $3.73 | $33.59 | $26.87 |
| 5 | | $26.87 | $2.69 | $24.19 | $19.35 |
| … | | | | | |
| 20 | | $0.19 | $0.02 | $0.18 | $0.14 |
| … | | | | | |
| **All Rounds** | | **$357** | **$36** | **$321** | **$257** |

$321 billion in disposable income) with an induced change in consumption of $257 billion. With the introduction of a 10% tax rate, the value of the multiplier fell to 3.57 from 5. With a proportional tax, the value of the multiplier can be calculated as

$$\text{Multiplier} = 1/(1 - \text{mpc} + t^*\text{mpc}),$$

where $t$ is the tax rate.

This also explains why a proportional tax system is an automatic stabilizer for the economy. The introduction of the proportional tax system lowered the value of the multiplier. Anytime there is an exogenous change in any type of spending (i.e., consumption, private investment, or government spending), the change in equilibrium income will be smaller with proportional taxes than without, helping stabilize the economy.

### "Fine-Tuning" the Economy

If the preceding example correctly represents how the economy actually works, then it appears rather easy for the government to use changes in government spending and changes in taxes to bring the economy to full employment. However, this is not the case. If the government could immediately (a) recognize that the economy is in recession, (b) agree to do something about the problem, (c) implement the program, and (d) get households and firms to respond to the plan, it might work. Unfortunately, each of these four steps takes time. The recognition, decision, implementation, and response lags take away valuable time during which the economy does not remain stagnant. It may be that by the time households respond to the plan, the economy will move from being in a recession. The policy taken to combat the recession will now be pointless or will create new problems for the economy. For this reason, fiscal policy is not used to "fine-tune" the economy—that is, keep the economy at full employment—but to combat persistent and significant macroeconomic imbalances.

## Temporary Versus Permanent Fiscal Policy Measures

Another problem when using fiscal policy to combat either a recession or inflationary pressure is that the government may want to make the policies temporary so that when the economy recovers, the discretionary policy measures do not create any unintended problems. Unfortunately, if a fiscal policy action is known to be temporary, it may not be effective. According to the permanent income hypothesis, first asserted by Milton Friedman (1957), consumers base their consumption spending on what they view as their permanent income. Because changes in income that households do not see as permanent will not lead to changes in consumption spending, temporary fiscal policy measures may have very small multipliers. A classic example of this problem is the Revenue and Expenditure Control Act of 1968, which created a temporary tax surcharge to help alleviate inflationary pressures in the economy. However, households did not see the increased tax as permanent, and it did little to lower expenditures or inflationary pressure. As Alan Blinder and Robert Solow (1974) explain, "Prices rose faster after its passage than before" (p. 10).

## Bias Toward Expansionary Fiscal Policy

The government can use expansionary fiscal policy, an increase in government spending or a reduction in taxes, to increase equilibrium income and contractionary fiscal

policy, a decrease in government spending and increase in taxes, to reduce equilibrium income and inflationary pressures. With the exception of worrying about any government deficits that may result, expansionary fiscal policy is a much more attractive policy in popular political terms than contractionary fiscal policy. Although it is unlikely that everyone will agree about which spending should be increased or which taxes should be cut, voters will be more likely to vote for an elected official who does either. On the other hand, few elected officials would look forward to a reelection after either decreasing government spending or increasing taxes. William Nordhaus (1975) explained the possibility of the existence of a political business cycle in which an elected official would choose more contractionary fiscal policies at the beginning of one's term but endorse more expansionary policies as the time for election draws new: "Moreover, with an incumbent's term in office there is a predictable pattern of policy, starting with relative austerity in early years and ending with the potlatch right before elections" (p. 187).

## Balanced Budget Rule

One of the dangers of expansionary fiscal policy is that it may increase the national debt. To finance a government budget deficit, the U.S. Treasury must borrow from either the public or the Federal Reserve by selling securities. A bias toward expansionary fiscal policy may lead to an ever-increasing national debt. To avoid this, there have been many calls for some type of balanced budget rule. However, whether the balanced budget rule is imposed through a constitutional amendment or some other mechanism, it would take away the ability of the government to use fiscal policy to counteract a recession and would also lead to greater instability in the economy (i.e., it will increase the value of the multiplier).

A balanced budget rule increases the value of the multiplier because it requires that when an exogenous decrease in spending lowers equilibrium income and threatens to bring the economy into recession, the government must reduce its own spending as tax revenues fall with income. Table 35.4 illustrates the particular example of a decrease of $100 billion in private investment spending, perhaps due to increased pessimism about the state of the economy among the business community. If the government is required to maintain a balanced budget, as private investment spending and income decreases, not only does this induce decreases in consumption spending, but tax revenues fall, which induce cuts in government spending to keep the government's budget balanced. At each round, income decreases because of decreases in both consumption expenditures and government expenditures. On net, the decrease in private investment spending will decrease equilibrium income by $556 billion; the multiplier is 5.56. With proportional taxes and a balanced budget rule, the multiplier is

$$\text{Multiplier} = 1/[(1 - \text{mpc})*(1 - t)].$$

Without a balance budget rule, this same $100 billion decrease in private investment would have decreased equilibrium income by only $357.

A major disadvantage of a balanced budget rule is the increase in the value of the multiplier. Anytime there is a major decrease (increase) in spending, not only can the government not use expansionary (contractionary) fiscal policy, but the resulting decrease in

**Table 35.4** The Balanced Budget Multiplier

| *Round* | *Initial Change in Spending* | *Change in Income* | *Change in Taxes* | *Change in Government Spending* | *Change in Disposable Income* | *Induced Change in Consumption* |
|---|---|---|---|---|---|---|
| 1 | $100.00 | $100.00 | $10.00 | $10.00 | $90.00 | $72.00 |
| 2 | | $82.00 | $8.20 | $8.20 | $73.80 | $59.04 |
| 3 | | $67.24 | $6.72 | $6.72 | $60.52 | $48.41 |
| 4 | | $55.14 | $5.51 | $5.51 | $49.62 | $39.70 |
| 5 | | $45.21 | $4.52 | $4.52 | $40.69 | $32.55 |
| … | | | | | | |
| 20 | | $2.30 | $0.23 | $0.23 | $2.07 | $1.66 |
| … | | | | | | |
| All Rounds | | $556 | $56 | $56 | $500 | $400 |

national income or inflationary pressure will be worse than if the government was not forced to respond to the change in spending.

## Full-Employment Budget Balance

Although the budget deficit worsens when the government uses expansionary fiscal policy and improves when it uses contractionary fiscal policy, the budget balance by itself may not be a good measure of whether the government is using expansionary fiscal policy. The budget deficit also changes with the business cycle. As an economy expands, tax revenues rise and government spending on social insurance programs falls; the government's deficit gets smaller. This is true whether or not the government is concerned that the economy's expansion may be excessive. So if the government's deficit is getting smaller while the economy expands, it is impossible to determine whether the decrease is due to automatic stabilizers or to discretionary fiscal policy just by looking at the net change in the budget balance. The same is true if the economy is heading into a recession. In that case, there would be a worsening of the budget deficit whether or not the government engages in expansionary fiscal policy. However, it is possible to calculate what the government's budget balance would be if the economy were at full employment. Also known as the *cyclically adjusted budget balance,* it adjusts the actual budget balance for any extra tax revenue the government would collect or social insurance payments it would save if the economy were not in a recession. With this control, if there is an increase in the full-employment budget balance, it must be because the government is engaging in contractionary fiscal policy while any decrease would indicate expansionary fiscal policy.

## The Problem of Crowding Out

Using expansionary fiscal policy may have some unintended consequences that can decrease the value of the multiplier. When the government increases its spending or decreases taxes, its budget balance declines. The only way the government can finance a deficit is to sell U.S. Treasury securities and, depending on who buys the securities, the increased deficit may lead either to increased inflationary pressure or higher interest rates.

If the government finances the deficit by selling U.S. Treasury securities to the Federal Reserve, this will increase the money supply and potentially increase inflationary pressure. Increases in the money supply will lower interest rates, encourage interest-sensitive consumption and private investment spending, and, if the economy is close to full employment, lead to increases in the price level.

On the other hand, if the government finances the deficit by selling U.S. Treasury securities to the public, this will put upward pressure on interest rates. This can be seen in the loanble funds market in Figure 35.5. The demand for loanable funds comes from the borrowing needs of businesses and government, and is inversely related to the interest rate. Because each possible private investment project available has an expected rate of return, businesses will take on those projects for which the expected rate of return exceeds the interest rate. As the interest rate falls, more projects are profitable, leading to an increase in the quantity of loanable funds demanded. This assumes that the government deficit, that is, demand for loanable funds on the part of the government, is not related to the interest rate but is determined by the size of the deficit. The supply of loanable funds reflects the relationship between saving and the interest rate in the economy. As the interest rate increases, the reward for saving increases, and there is an increase in the quantity supplied of loanable funds. In the figure, D shows the demand for loanable funds before the expansionary fiscal policy, and D' is the demand for loanable funds with the increased borrowing needs of $100 billion from the government. Prior to the government engaging in expansionary fiscal policy and increasing its demand for loanable funds, the equilibrium interest rate is 5% and the equilibrium level of loanable funds is $500 billion. After the government borrows $100 billion to cover its deficit, the equilibrium interest rate rises to 5.5% and the equilibrium level of loanable funds rises to $550 billion. Because the government's share of loanable funds rose by $100 billion, but the total amount of loanable funds rose by only $50 billion, the increase in the interest rate must have decreased the quantity demanded of loanable funds by $50 billion. In other words, the increased borrowing needs of the government crowded out private investment.

Additionally, the increase in interest rates may affect the exchange rate leading to changes in the trade balance

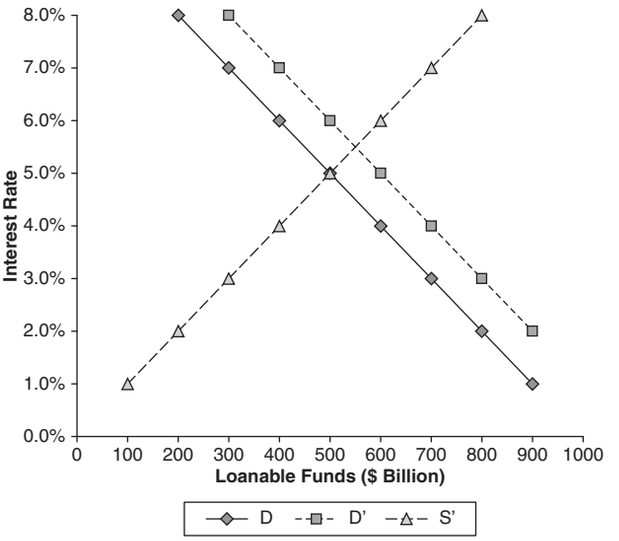


**Figure 35.5** The Market for Loanable Funds

that further reduce the effectiveness of expansionary fiscal policy. As interest rates in the United States increase, both domestic residents and foreigners will want to buy U.S. assets. In the market for the U.S. dollar, this will increase the demand for dollars (foreigners wanting to buy U.S. assets first need to buy U.S. dollars) while decreasing the supply (Americans who may have bought foreign assets and consequently foreign currencies will not do so). Both factors will lead to an appreciation of the dollar and encourage an increase in U.S. imports of foreign goods and services while discouraging foreign purchases of U.S. exports. This decrease in expenditure on domestic goods and services will further limit the increase in equilibrium income in response to the expansionary fiscal policy.

It is even possible that there may be complete crowding out of private investments. Given the effects on the interest rate and the exchange rate, the increase in government spending might just provoke a decrease in spending on domestic goods and services that completely offsets any increase in national income.

## Ricardian Equivalence

An alternative view of why expansionary fiscal policy may not increase national income is the Ricardian equivalence theorem. As Robert Barro (1992) explains it, the demand for goods by households is a function of the expected present value of taxes. Households determine how much to consume based on their net wealth position or the expected present value of income less the expected present value of taxes. Any increase in government spending or decrease in taxes financed through an increase in the deficit will only lead to an increase in expected future taxes equal to the same present value as the fiscal policy stimulus and will not lead to an increase in consumption spending. In Barro's (1989) words, "Fiscal policy would affect aggregate consumer demand only if it altered the expected present value of taxes" (p. 39).

### Empirical Evidence

Although the advocacy of fiscal policy to affect the economy began with John Maynard Keynes's publication of *The General Theory of Employment, Interest and Money,* it was only when the government increased military spending for World War II that his theory was truly implemented. During the Great Depression, Franklin Roosevelt experimented with Keynesian theory when he implemented the New Deal. Although the programs of the New Deal were implemented prior to the publication of *General Theory,* its emphasis on increased government spending to stimulate the economy is exactly what Keynes was advocating.

After World War II, Keynesian economics and the beneficial effects of fiscal policy were definitely in vogue. Through the 1950s and 1960s, the economy behaved in a manner similar to the Keynesian model. The economy varied between times of low unemployment and high inflation and of high unemployment and low inflation. The Keynesian policy prescription was clear: Use expansionary fiscal policy to cure an unemployment problem and a contractionary one to cure inflation.

## Emphasis on Monetary Policy

Unfortunately, in the 1970s, the rapid increase in the price of oil appeared to change the relationship between unemployment and inflation. During this time, the economy experienced *stagflation,* that is, the combination of both unemployment and inflation. The Keynesian model seemed to be failing to explain the economy, and fiscal policy did not seem to hold much potential to mitigate the problems. At the same time, increases in the size of the U.S. government deficit and national debt caused many to worry about the effects of further expansionary fiscal policy on future generations. According to John Taylor (2000), by the end of the 1970s, discretionary countercyclical fiscal policy was no longer considered a serious policy option in the United States. As Laurence Seidman (2001) explains, "Whereas Congress enacted a major fiscal stimulus package to counter the 1975 recession, it did not even consider a serious package in the 1991 recession" (p. 18). The focus had shifted to the use of monetary policy. As Taylor (2000) writes, "Over the [1980s and 1990s], the Federal Reserve's interest rate decisions have become more explicit, more systematic, and more reactive to change in both inflation and output" (p. 21). The economy performed well under this regime with only two relatively mild recessions, one in 1980 to 1982 and one in 1990 to 1991.

However, signs of the weakness of monetary policy became apparent when looking at the slump in the Japanese economy. After amazing postwar growth that propelled war-torn Japan to a major industrial nation, a real estate and stock market boom in the 1980s was followed by a crash in both markets and a period of relatively slow growth or recession since 1991. This period has also been characterized by very low interest rates as the Bank of Japan tried to stimulate the economy with expansionary monetary policy. Similar to our experience during the Great Depression, the Japanese economy had fallen into a liquidity trap. First described by Keynes, a *liquidity trap* is when interest rates are at or close to zero but the low interest rates do not spur interest-sensitive spending. Because an expansionary monetary policy affects the economy when an increase in the money supply lowers interest rates and encourages interest-sensitive spending, when interest

rates are at or close to zero, there is little impact on the economy. As in the Great Depression with the impotence of traditional monetary policy, Japan turned to fiscal policy to help promote growth in the economy.

### Economic and Financial Crisis of 2008

After a relatively long expansion and large increases in stock price during the late 1990s, the economy began to show signs of weakness in early 2000. What has been called the *dot-com bubble,* the large increase in the stock prices of technology stocks, burst in early 2000, and for most of 2001, the economy was in recession. The recession, along with the terrorist attacks of September 11, 2001, led the Federal Reserve to lower interest rates significantly. The Fed lowered the target federal funds rate 11 times in 2001 from a rate of 6.00% to 1.75%. The decrease in interest rates helped lead the economy out of recession in part by increasing the demand for new homes. Housing prices soared along with consumer spending in response to higher wealth and to lower interest rates. From the beginning of 2001 until housing prices peaked in mid-2006, according to the S&P/Case-Shiller Price Index, housing prices essentially doubled in 10 major metropolitan cities. However, once housing prices began to fall and the subprime mortgage crisis ensued, the economy began to slow and went into recession in December 2007. Although the Federal Reserve attempted to encourage consumption and private investment spending through further decreases in interest rates, by the end of 2008, the target rate for the federal funds rate was essentially zero, and the U.S. economy faced its own liquidity trap. Although the Federal Reserve continued to provide liquidity into the markets through various lending facilities and quantitative easing, policy makers began looking once again to fiscal policy to help the economy out of recession.

### Size of the Multipliers

A major controversy about using fiscal policy concerns the size of the government spending and tax multipliers. As previously discussed, the multipliers depend on the magnitude of the marginal propensity to consume and the tax rate. However, it is not so simple to calculate the size of the multipliers in the real world. E. Cary Brown (1956) was one of the first to attempt to estimate the sizes of the multipliers by looking at the effectiveness of fiscal stimulus during the Great Depression. Using Brown's work, Gauti Eggertsson (2006) measured the real spending (holding the budget deficit constant) multiplier to be 0.5 and the deficit spending (cutting taxes and accumulating debt while holding real government spending constant) multiplier as 2 during the 1930s.

More recent estimates on the values of the multiplier vary widely and are the topic of great debate within the economics profession. Olivier Blanchard and Roberto Perotti (2002) estimate that the value of the multipliers are close to 1. They explain, "While private consumption increases following spending shocks, private investment is crowded out to a considerable extent. Both exports and imports also fall." On the other hand, Eggertsson (2006) finds that "each dollar of real government spending increases output by 3.37 dollars and each dollar of deficit spending increases output by 3.76 dollars. These multipliers are much bigger than have been found in the traditional Keynesian literature." Christina Romer, President Obama's Chair of the Council of Economic Advisors, and Jared Bernstein, chief economic advisor to Vice President Biden, arguing in support of the American Recovery and Reinvestment Plan, estimate the output effects of a permanent increase of government purchases of 1% of GDP to be 1.57% after 10 quarters, and of a permanent tax cut of 1% of GDP to be 0.99%.

## Future Directions

A big challenge facing fiscal policy in the future is meeting the increasing costs of Social Security, Medicare, and Medicaid programs. Penner and Steurele (2007) estimated that between 2006 and 2010, the cost of these three programs would grow by $326 billion, while the growth in tax revenues is expected to be only $494 billion. As they explain, "The three programs will absorb two-thirds of all revenue growth, even though they still only constitute[d] about 40 percent of total spending in 2006" (p. 3). Barring a broad reform of these programs, Penner and Steurele argue that some type of "trigger" system for these benefits be constructed. For example, a trigger system would set a particular growth rate for these benefits when the economy is close to full employment and another during a recessionary period.

Seidman (2001) argues for improving existing automatic stabilizers. As he asks, "Why not automatically trigger a tax cut in a recession?" (p. 39). By this, Seidman means a tax cut in addition to the automatic fall in revenue due to our progressive tax structure. Such a trigger would not require any legislative action but therefore would occur much more quickly than under the current system.

Another important discussion concerns the highly political atmosphere in which fiscal policy is set. Having served on both the Council of Economic Advisors and the Board of Governors of the Federal Reserve System, Blinder (1997) describes his different experiences with the setting of fiscal and monetary policies. Whereas monetary policy is in the hands of the Federal Reserve, an agency relatively independent of politics, fiscal policy is the result of a highly political process. As Blinder explains, although a policy discussion begins with social welfare as its main goal, it often "turns to such cosmic questions as whether the chair of the

relevant congressional subcommittee would support the policy, which interest groups would be for and against it, what that 'message' would be, and how that message would play in Peoria." *The Economist* ("Fiscal Flexibility," 1999) has called for serious consideration of an independent fiscal authority. Seidman (2001) proposed that members of such a group would be appointed in a similar manner as the members of the Federal Reserve. The budget would continue to be set by the president and Congress, and "the board would implement periodic adjustments in taxes and government spending relative to the budget enacted by Congress."

## Conclusion

Fiscal policy has been and continues to be an important stabilization policy of the government. This chapter reviews the theory of fiscal policy and how it has been used in the United States. It gives insight into some of the major debates surrounding fiscal policy and future directions for research. Although the efficacy of discretionary policy measures continues to be debated and there is concern about the growing expenditures for Social Security, Medicare, and Medicaid, government spending and taxes will continue to be an important component of the policy toolbox available to influence the economy.

## References and Further Readings


Anderson, J. (2006, May 22). Tax misery and reform index. *Forbes.* Retrieved January 29, 2009, from http://www.forbes.com/global/2006/0522/032.html

Auerbach, A. J. (2007). American fiscal policy in the post-war era: An interpretive history. In B. Eichengreen, D. Stiefel, & M. Landesmann (Eds.), *The European economy in an American mirror* (pp. 214–232). New York: Routledge.

Barro, R. J. (1989). The Ricardian approach to budget deficits. *Journal of Economic Perspectives, 3*(2), 37–54.

Bayoumi, T., & Sgherri, S. (2006, July). *Mr. Ricardo's great adventure: Estimating fiscal multipliers in a truly intertemporal model* (IMF Working Paper). Retrieved January 29, 2009, from http://ssrn.com/abstract=920260

Blanchard, O., & Perotti, R. (2002). An empirical characterization of the dynamic effect of changes in government spending and taxes on output. *Quarterly Journal of Economics, 117*(4), 1329–1368.

Blinder, A. S. (1997). Is government too political? *Foreign Affairs, 76*(6), 115–126.

Blinder, A. S., & Solow, R. M. (1974). Analytical foundations of public policy. In S. Blinder, R. M. Solow, G. F. Break, & P. O. Steiner (Contributors), *The economics of public finance* (pp. 3–118). Washington, DC: Brookings Institution.

Brown, E. C. (1956). Fiscal policy in the 'thirties: A reappraisal. *American Economic Review, 46*(3), 857–879.

Congressional Budget Office. (2008, December 23). *Historical effective tax rates, 1979 to 2005: Supplement with additional data on sources of income and high-income households.* Retrieved January 29, 2009, from http://www.cbo.gov/doc.cfm?index=9884

Congressional Budget Office. (2009, January 7). *Revenues, outlays, surpluses, deficits, and debt held by the public, 1968 to 2007.* Retrieved January 29, 2009, from http://www.cbo.gov/budget/data/historical.xls

*Economic report of the president. Transmitted to the Congress February 2008, together with the annual report of the Council of Economic Advisers.* (2008). Washington, DC: Government Printing Office. Retrieved January 10, 2009, from http://www.gpoaccess.gov/eop

Eggertsson, G. B. (2006). *Fiscal multipliers and policy coordination* (Federal Reserve Bank of New York Staff Reports No. 241).

Fiscal flexibility. (1999). *The Economist, 353*(8147), 80.

Friedman, M. (1957). *A theory of the consumption function: A study by the National Bureau of Economic Research.* Princeton, NJ: Princeton University Press.

Keynes, J. M. (1936). *The general theory of employment, interest and money.* New York: Harcourt, Brace.

Krugman, P. (1998). *It's baaack: Japan's slump and the return of the liquidity trap* (Brookings Papers on Economic Activity No. 29). Retrieved January 29, 2009, from http://www.brookings.edu/press/Journals/1998/bpea198.aspx

Micl, T. (2008). Rethinking fiscal policy. *Challenge, 51*(6), 91–104.

Nordhaus, W. D. (1975). The political business cycle. *Review of Economic Studies, 42*(2), 169–190.

Penner, R. G., & Steuerle, C. E. (2007, August 1). Stabilizing future fiscal policy: It's time to pull the trigger. *The Urban Institute.* Retrieved January 29, 2009, from http://www.urban.org/publications/411524.html

Romer, C., & Bernstein, J. (2009, January 9). *The job impact of the American Recovery and Reinvestment Plan.* Retrieved January 29, 2009, from http://otrans.3cdn.net/ee40602f9a7d8172b8_ozm6bt50i.pdf

Schultze, C. L. (1992). Is there a bias toward excess in U.S. government budgets or deficits? *Journal of Economic Perspectives, 6*(2), 25–43.

Seidman, L. (2001). Reviving fiscal policy. *Challenge, 44*(3), 17–42.

Spencer, R. W., & Yohe, W. P. (1970, October). The "crowding out" of private expenditures by fiscal policy actions. *Federal Reserve Bank of St. Louis Review,* pp. 12–24.

Standard & Poor's. (2009). *Case-Shiller home price history.* Retrieved January 20, 2009, from http://www2.standardandpoors.com/spf/pdf/index/CSHomePrice_History_123062.xls

Taylor, J. B. (2000). Reassessing discretionary fiscal policy. *Journal of Economic Perspectives, 14*(3), 21–36.

U.S. Census Bureau. (2008, July 1). *State and local government finances by level of government and by state: 2005–06.* Retrieved January 29, 2009, from http://www.census.gov/govs/estimate/0600ussl_1.html

# 36

# Government Budgets, Debt, and Deficits

Benjamin Russo

*University of North Carolina at Charlotte*

This chapter describes the economics of government budgets, with particular attention to government borrowing in the United States at the federal, state, and local levels. The next section provides definitions and describes some principles of federal, state, and local budgeting. The third section focuses on major issues pertaining to federal government borrowing. The final section is a summary.

## Definitions and Principles of Government Budgeting

### Preliminaries

Equity and efficiency are crucial concepts in economics, so it is useful to begin by defining these terms. Two principles guide evaluations of the effect economic policies have on equity. The *benefit principle* is the viewpoint that it is equitable for citizens who benefit from government services to pay for them. The *ability-to-pay principle* is the viewpoint that it is equitable for a citizen's tax liabilities to be correlated with the citizen's economic resources. It follows that taxpayers with equal abilities to pay should be taxed equally (horizontal equity) and that tax liability should increase as ability to pay increases (vertical equity). Resources are allocated efficiently if and only if they are used where they have the highest social value.

### Operating and Capital Budgets

State and local governments report operating budgets and capital budgets. In general terms, an *operating budget* is an enumeration of (a) expenditures on current operations and (b) the revenue inflows required to finance those operations (current operations take place each period). Typical expenditures on current operations include purchases of services and of tangible items. *Services* are commodities that cannot be stored, for example, state employee salaries and interest payments on government debt. *Tangibles* are items that are used up each period, for example, stationery and gasoline. Revenue inflows collected in 2009 include tax and fee collections that are used to pay for 2009 operating expenditures.

*Capital budgets* enumerate (a) planned spending on purchases and repairs of capital assets and (b) the means of financing capital assets. In contrast to current operations, capital assets are used for many periods (examples are roads, school buildings, and water treatment plants). Because capital purchased in the current period produces services in the future, capital, like houses, often is financed with borrowed funds. Governments borrow by selling bonds, which are repaid over many periods, when the services produced by capital are enjoyed. A legal obligation requiring future expenditures is a financial liability. Thus, a capital budget is an enumeration of expenditures on capital assets and newly incurred liabilities in a given period. In contrast to most state and local

governments, the federal government does not report a capital budget.

A fiscal year is a 12-month period covered by an operating budget. The U.S. federal government's fiscal year begins on October 1 of each calendar year and ends on September 30 of the following calendar year. Before the beginning of a fiscal year, governments formulate *planned* operating budgets, which show planned expenditures and expected tax and fee revenues. However, some operating expenditures are difficult to plan for (e.g., disaster assistance), and tax and fee revenues are difficult to predict (e.g., sales tax revenue). Thus, within any fiscal year, actual expenditures may differ from plans, and actual tax and fee revenue may fall short of predictions. In all cases except Vermont, U.S. state governments are constitutionally required to adjust their operating budgets so that by the end of each fiscal year, actual expenditures do not exceed tax and fee revenue. Economists call this restriction the government's *current budget constraint.* Note that if expenditures equal tax and fee revenue, the state budget is said to be balanced.

Although the national government also must satisfy a current budget constraint, the restriction is much less severe than that faced by subnational governments. This is true because national governments, generally, are not required to balance their operating budgets. Thus, national governments can borrow to finance operating expenditures. The national current budget constraint states simply that in a fiscal year, government expenditures cannot exceed the sum of borrowed funds plus tax and fee revenue. Additionally, many national governments can issue money, which can be used to finance expenditures. Of course, there is a limit to money finance because money creation tends to cause inflation. U.S. law precludes the federal government from directly issuing money to finance expenditures. However, if the U.S. Federal Reserve Bank (the Fed) buys U.S. Treasury bonds, the money stock increases. This is a form of indirect money finance. However, U.S. law precludes the federal government from compelling the Fed to buy Treasury bonds. For most of its history the Fed appears not to have engaged in indirect money finance.

Governments' current budget constraints restrict expenditure and revenue choices within a fiscal year. Governments, like individual consumers and businesses, also face a constraint that extends into future periods, namely, the intertemporal budget constraint (IBC). IBC dictates that the sum of current and all future expenditures plus currently outstanding debt cannot exceed current plus all future tax and fee revenue (*outstanding debt* is debt that has not yet been paid off). Here, future expenditures/revenues are measured in "present value terms." (Note that the present value of a dollar to be received in a future period is the dollar adjusted downward by the opportunity cost incurred by not having the dollar for a period. The opportunity cost is the rate of return that could have been earned if the dollar were invested during the period.)

There seems to be a great deal of public confusion about IBC, so it is important to be clear about what restrictions it does and does not impose. In general, the IBC imposes the condition that any increase in expenditure implies an increase in current taxes or future tax liabilities equal in present value to the expenditure. By way of a false analogy, IBC sometimes is taken to mean that governments must repay principal on their debts. This is incorrect. To see why, first note the economic incentives that motivate creditors to lend. Creditors are willing to forgo consuming their wealth for some period of time if they are *compensated with a market rate of return* on the loans. Second, because individual consumers, sadly, live only finite periods of time, they must repay debt principal before they or their estates expire. Creditors will demand that the principal is repaid. For if creditors were not repaid the principal, they would suffer a large opportunity cost, namely, the opportunity to continue to earn the market rate of return.

However, the same cannot be said for governments and successful business firms because their "lifespans" effectively extend into the indefinite future: They are not finite. As a result, if a government reliably pays the market rate of return on its outstanding debt, and creditors believe the government will continue to do so, the government can satisfy its creditors by servicing its debts forever; it is not required to repay the principal on the debt.

IBC may seem to be a weak restriction on government choices. Nonetheless, IBC does place an important restriction on government debt policy: The government cannot sustain a policy that would cause debt to grow faster than the economy in the long run. If the debt did grow faster, it would outrun the economy's ability to generate the revenue that must be collected to pay the market rate of interest on the debt. In this case, creditors eventually would stop making loans to the government, and the policy must terminate. Is recent U.S. debt policy sustainable?

## Why Report Budgets?

The U.S. government reports a budget each fiscal year in accord with Section 9, Article 1, of the U.S. Constitution, which states that "a regular statement of accounting of receipts and expenditures of all public money shall be published from time to time." The states have stringent constitutional provisions, requiring construction of budgets, and states impose similar requirements on their local governments.

But what is the value added of a government budget? Government budgets have two major purposes. First, budgets provide information the public needs to evaluate the costs and benefits of public goods and services. This information is fundamental in democracies. Underlying the relations between a democratic government and its citizens is an assumption that government activities are sanctioned by a social contract in which citizens relinquish some freedom in return for public goods and services. Without a budget, this evaluation would be more difficult than it is.

Second, budgets indicate whether governments' expenditure and revenue decisions are financially responsible. Without a budget, it is harder to know if fiscal policy satisfies the government's budget constraints, and it would be difficult, if not impossible, for the public and the government itself to know whether spending, revenue, and borrowing policies are financially sound.

## Why Report Capital Budgets?

State and local governments separate operating from capital expenditures, and report a separate capital budget. For example, spending on public capital infrastructure (government buildings, highways and bridges, etc.) is included in state capital budgets, whereas spending on teacher and administrative salaries is included in operating budgets. Many foreign national governments publish capital, as well as operating, budgets. Capital budgeting requires time and resources.

Do capital budgets have economic benefits that offset these costs? Public capital goods generate returns and impose risks, just as private capital does. Governments issue bonds, so there is risk of government bond default. Governments also maintain portfolios in trust for the public (largely in the form of pensions). There is a risk to taxpayers that pension assets will lose value. Government capital budgets allow taxpayers and investors to more easily evaluate the risk and returns of funds held in trust by the government.

State and local governments are required to balance their budgets. If there were no separation of capital expenditures from current expenditures, the balanced budget requirements would prevent state and local governments from borrowing to finance capital assets. In this case, states would need to finance capital expenditures by collecting taxes in the period the purchases were made: In no period could capital purchases exceed tax revenue collected in that period. This would be inefficient and inequitable.

To see why it would be inefficient, note that financing capital purchases from current tax collections would require outsized tax increases in periods when purchases are made, because capital, by nature, is very costly relative to a single period's income. Tax finance of capital would lead to large variations in tax revenue requirements because purchases of capital tend to be "lumpy." For example, once a new school or courthouse is built, government does not need to build another for a while. As explained next, tax variability is economically inefficient (Barro, 1979). As well, the expenditure could require a single-period increase in tax revenue too large to be politically tenable, which could lead to a less than efficient level of investment in public capital.

Financing capital purchases from current tax collections would also be inequitable. First, it would violate the benefit principle because future generations could consume government capital services they have not paid for. Borrowing today and repaying part of the debt with future tax dollars allows the cost to be shared with the future beneficiaries of capital services. Second, it would violate horizontal equity because households with equal abilities to pay, and who receive equal government services but are born to different generations, would face different tax burdens.

## Should the U.S. Government Report a Capital Budget?

The U.S. federal government does not publish a capital budget. Some economists argue that publication of a federal capital budget would be beneficial. Others argue that a federal capital budget would be difficult to implement.

A federal capital budget would make it easier for taxpayers and investors to evaluate the risk and returns of federal government investments. Marco Bassetto and Thomas Sargent (2006) argue that the absence of a federal capital budget may tempt some policy makers to make choices inconsistent with Generally Accepted Accounting Principles. For example, it is sometimes argued that the government should sell some of its physical assets to reduce the federal government's debt. In general, this is not an economically meaningful way of satisfying the government's budget constraint; instead, it merely replaces a physical asset for cash, and has no effect on social net wealth. The exchange would be economically meaningful only if the private sector pays more for the asset than its true social value. This seems unlikely in most cases. If the federal government published a capital budget, the economic futility of such an exchange would be easier to understand and avoid.

It has been argued that it would be hard to implement a federal capital budget because it is difficult to categorize some public expenditures. Examples are highway spending, public education, and spending on public safety. State governments address this issue by including the cost of school building in capital budgets and teacher salaries in operating budgets. It is unclear why the federal government could not or should not take the same approach.

## Weaknesses in Government Budgeting

Budgets should be designed to aid policymaking and provide the public with information about fiscal policy. In this way, budgets provide rough guides to the social costs and benefits of government. However, these goals can be subverted in a number of ways.

First, the government can impose costs and provide benefits in ways that are not captured in budgets. For example, the U.S. federal government has imposed unfunded mandates on subnational governments (e.g., requirements for school achievement), but the costs of such programs do not appear in the federal budget.

Second, many government-imposed costs and benefits are intangible and hard to measure. For example, it may be impossible to measure the social benefits of public

education. Even if possible, the cost of conducting the measurements could be prohibitive. In the case of government guarantees of private loans, an implicit social cost is created because taxes may have to be increased if the loans go into default; however, the social cost is not reported in budgets. As well, governments can rule some expenditures "off budget." In this case, although measurements are available, the government does not report them in its operating budget. At the federal level, Social Security is an example of an off-budget program.

Third, government budgets may be constructed using substandard accounting. For example, most governments use *cash flow accounting.* Under cash flow accounting, revenues are included in the budget in the period when received: Expenditures are included in the period when they are paid. In contrast, Generally Accepted Accounting Principles, which are established by the Financial Accounting Standards Board and must be followed by all publicly traded corporations, require private firms to use *accrual accounting.* In accrual accounting, revenues are included in the budget in the period earned, and costs are included when incurred. Cash accounting is easier (than accrual accounting) to manipulate in ways that produce a misleading representation of the government's true fiscal position. For example, if a state's operating budget is out of balance, the state may postpone payment for goods and services until the following fiscal year, or may order tax liabilities due in the following year be prepaid, making the budget to appear to be in balance. Cash flow accounting manipulation may delay the appearance of budget problems but does not solve them. In the interim the government's fiscal position can worsen.

In the United States, delay and erosion in the federal government's fiscal position appear to be most severe in the Social Security and Medicare programs. Social Security uses "pay-as-you-go" finance—that is, currently employed workers pay Social Security (payroll) taxes, and the government uses the tax collections to make benefit payments to current beneficiaries (mostly retirees). However, pay-as-you-go Social Security satisfies the government's intertemporal budget constraint only if there are sufficient young taxpaying workers to finance benefits. Assuming the system initially is in balance, if the number of young workers subsequently declines relative to the number of beneficiaries, the system will no longer be sustainable: To satisfy the constraint, benefits must be reduced, taxes must be increased, or both. In the past 35 years, the average U.S. fertility rate has declined, diminishing the number of young workers per retiree, and the mortality rate of retirees has decreased, increasing the number of retirees per young worker. As currently structured, the U.S. Social Security system appears unsustainable, but this is not reflected in the budget.

In any year, the lion's share of U.S. federal and state budgets consist of previously established programs. The majority of budgeted programs are not subjected to systematic review of their costs and benefits. Programs that outlive their original motivation and usefulness can become a net burden on society, subverting a basic reason for budgeting. Some observers advocate *zero-based budgeting.* In this case, the government would reevaluate each program at the beginning of each budget year. Such close monitoring of government programs has a potential to reduce unproductive government spending. However, close monitoring imposes administrative costs. Zero-based budgeting would be an improvement if the cost of unproductive programs exceeds the administrative costs of monitoring programs.

## State and Local Government Budget Shortfalls: Rainy Day Funds

From time to time, actual expenditures expand faster than tax and fee revenue, or revenue may actually decline, causing a budget shortfall. Budget shortfalls often occur during economic contractions. Shortfalls are created as the decline of household and business income reduces income tax revenue, consumer spending, and sales tax revenue. Occasionally, property values also decline, reducing property tax revenue. Making matters worse, during contractions, demand for government expenditures on social programs tend to increase (e.g., unemployment insurance and social welfare spending increase in recessions).

Can budget shortfalls be avoided by eliminating the instigating factor, economic contraction? The National Bureau of Economic Research defines the *business cycle* as the recurring and persistent contractions and expansions in economic activity. The term *recurring* is emphasized because past experience offers convincing evidence that economic contractions are indicative of developed economies and will recur from time to time. The correlated budget shortfalls create enormous fiscal stress, because states must balance their operating budgets. The question here is, how can state and local governments respond to periodically recurring budget shortfalls?

One response is to increase taxes and reduce expenditures. This approach can be counterproductive and tends to be politically difficult. It can be counterproductive because reducing demand for goods and services can worsen the shortfall. It is difficult to increase taxes in contractions because income declines as workers lose jobs and firms go out of business. However, there is a sound alternative to tax increases and expenditure cuts during contractions. Because recurring contractions are almost a certainty, sound fiscal planning would have state and local governments act before economic disaster strikes by establishing budget stabilization funds, also known as *rainy day funds* (RDFs). In the ideal case, during economic expansions, RDFs would accumulate tax revenue in trust funds. During subsequent contractions, the trust funds would be used to support expenditures, reducing the need to raise taxes and cut expenditures.

The Advisory Council on Intergovernmental Relations, a group of tax administrators and state government officials,

recommends that RDFs be established at a target level of 5% of a state's annual expenditures. However, the 5% target has proven to be very inadequate. For example, in recessions in 1981 and 1982, 1990 and 1991, and 2001, North Carolina budget shortfalls averaged 19% of a typical year's expenditure. There is evidence that absent or insufficient RDFs contribute to fiscal stress, imposing substantial yet avoidable costs on society (Russo, 2002). How?

First, without a sufficient RDF, policy makers and administrators spend substantial time and energy reworking policies in response to budget shortfalls, rather than on the formulation of social policy and the ordinary business of running the government. Ignoring the ordinary business of government causes government services to deteriorate.

Second, without a sufficient RDF, state and local governments risk lower credit ratings from bond rating agencies. Lower credit ratings cause interest rates on state and local borrowing to increase, which leads to avoidable increases in future tax liabilities.

Third, without a sufficient RDF, there is a risk to private investment and planning because taxes would otherwise need to be increased, or expenditures cut, or both, during economic contractions. Stop-and-go fiscal policy increases uncertainty, which discourages private investment and tends to retard economic growth.

Fourth, without a sufficient RDF, taxes tend to be more variable than otherwise. More tax variability implies a larger excess burden of taxation. The *excess burden* of a tax is the social burden of the tax in excess of the amount of tax revenue collected. Excess burden occurs because people substitute lower valued untaxed commodities for higher valued taxed commodities, reducing economic welfare by more than the tax payment remitted to the government. Without RDFs, policy makers tend to respond to budget shortfalls by increasing tax rates. An appropriately sized RDF can reduce excess burden by reducing the need to increase tax revenue during budget shortfalls.

## National Debt and Deficits

### What Is the "National Debt"?

National government financial debt is created when the expenditures in a fiscal year exceed tax and fee revenue. The national government finances the excess of expenditure over revenue by borrowing. For example, the U.S. government auctions off U.S. Treasury bills, notes, and bonds (U.S. bonds, hereafter). The U.S. Treasury defines the *Gross Debt* to be the value of all national government bonds outstanding. However, the media and the public often refer to this same concept as the *National Debt.* Gross Debt and National Debt are synonyms. The remainder of this chapter uses the former term.

A government's ability to sell bonds depends on creditors' faith that the bonds will be repaid and on the creditors' financial capacity to buy the bonds. Creditors' faith depends on the economy's ability to generate future tax revenue, which depends on the economic growth rate. Creditors' financial capacity depends on their incomes and saving. Because saving cannot grow faster than the economy in the long run, creditors' financial capacity to buy bonds is limited by the economy's long-run growth rate. Therefore, if Gross Debt grows faster than the economy in the long run, the economy's ability to generate tax revenue to service debt, and creditor's capacity to buy new debt, falls short of the government's liabilities. If the public understands that the Gross Debt continually grows faster than the economy, creditors will demand abnormally high interest rates to compensate for the risk of not being repaid. Or creditors may simply refuse to purchase additional government debt. In this case, the government faces a "debt crisis," and the government's choices become severely constrained: (a) taxes must increase or expenditure must be cut, or both, or (b) government must default on its liabilities and convince creditors to reschedule debt payments. If these options are not available, the only recourse is for the government to repudiate the debt. In this case, government simply declares its intention to not repay the debt.

Interest rates tend to increase greatly in debt crises, and living standards tend to decline. If the government responds to a debt crisis by raising taxes and reducing spending, this exacerbates the decline. Fortunately for the United States, its national government has never experienced a debt crisis. Nevertheless, some public finance economists (Blocker, Kotlikoff, & Ross, 2008; Gokhale, Page, Potter, & Sturrock, 2000; Kotlikoff, 1992) worry that in the decades ahead U.S. Medicare and Social Security obligations could generate fiscal imbalances that could approximate a debt crisis. They argue that U.S. tax rates must increase substantially or benefits must be cut by painful amounts.

Countries that default on debt often have attempted to maintain financial viability by rescheduling debt payments. In this case, the government usually ends up paying much higher interest rates to compensate for the increase in risk perceived by creditors. Even though interest rates increase, the exchange value of the country's currency tends to decline greatly, as creditors shift funds to safer environments. A large decline in the currency reduces the standard of living because imports become more expensive.

### Debt Repudiation

Debt repudiation is the most drastic response to a fiscal crisis. One form of debt repudiation occurs when a government renounces its debt obligations. In 1917, Russia announced that it would not repay debt incurred by the Czars. In the 1930s, the Peronists renounced Argentina's government debt. A second form of debt repudiation occurs when a government prints money to repay outstanding debt. This form of repudiation often leads to inflation that reduces the real value of debt. After World War I, the Allies imposed war reparation payments on Germany. The

German government created a hyperinflation that effectively made the reparation payments worthless.

An advantage of debt repudiation is that it eliminates the burden on taxpayers of repaying previously issued debt. A disadvantage is that repudiation usually decimates the value of a country's currency, its ability to trade, and the standard of living. For a long time after repudiation, a country may not be able to borrow at all, making it difficult to maintain government services.

## How Large Is the U.S. Gross Debt?

The U.S. Gross Debt was about $10 trillion at the end of the federal government's 2008 fiscal year. As of early 2009, the U.S. Congressional Budget Office estimated that fiscal 2009 borrowing would increase the Gross Debt by about $1.2 trillion.

However, this figure could provide a misleading measure of the effects the debt has on the U.S. standard of living. There are a number of reasons for this.

First, the government's ability to service its debt, as well as the extent of the debt's effect on interest rates and the standard of living, depend on the size of the economy. For example, a debt of $10 trillion would overwhelm a small country such as Costa Rica, whose 2007 gross domestic product (GDP) is estimated to be about $48 billion (CIA World Factbook, 2008). If Costa Rica's debt were this large, interest rates would rise to extreme levels, and the value of its currency probably would decline precipitously along with its standard of living. The $10 trillion U.S. Gross Debt does not cause these dire results because U.S. GDP, at a little less than $14 trillion in 2008, is 287 times larger than Costa Rica's. Thus, the U.S. debt is a much smaller fraction of its credit markets, so the debt is much easier to service and sustain. As a result, the absolute amount of debt in countries cannot usefully be compared unless each country's debt is adjusted for the scale of the economy. In fact, a single country's debt in different periods cannot be usefully compared because the scale of the economy increases over time.

Because economic scale differs across countries and over time, a more accurate assessment of the economic impact of debt is made by adjusting the absolute amount of debt for scale. A common way to do this is to divide the absolute amount of a country's national debt by its GDP. When the U.S. Gross Debt was $10 trillion, the debt/GDP ratio was about 71% ($10 trillion/$14 trillion). In historical comparison, the U.S. debt/GDP ratio was 120% at the end of World War II (WWII) and did not decline to 73% until the mid 1950s. During the Clinton administration, the ratio declined to about 60% (see Figure 36.1). The 2007, debt/GDP ratios in seven large capitalist-oriented economies were 64.2% in Canada, 104% in Italy, 170% in Japan, 63.9% in France, 64.9% in Germany, and 43.6 % in the United Kingdom. The arithmetic average is 85.1%.

Second, the U.S. Federal Reserve holds a substantial fraction of the Gross Debt. Federal Reserve holdings of the debt do not impose the same tax burden as other publically held debt because the Fed remits interest income it earns on U.S. bonds back to the Treasury. Prior to the 2008 financial crisis, the Federal Reserve held about $800 billion of the Gross Debt. Subtracting this amount from the Gross Debt reduces the debt/GDP ratio to about 65.7% ($9.2 trillion/$14 trillion).

Third, federal agencies, such as the Social Security Administration, have large holdings of federal government bonds. At the end of fiscal 2008, intra-agency debt was more than $4 trillion. Some public finance economists argue that this debt is not a true burden on taxpayers: Although it is a liability of one government agency, it is an asset to another agency. However, this point is controversial, because Social Security and Medicare suffer from

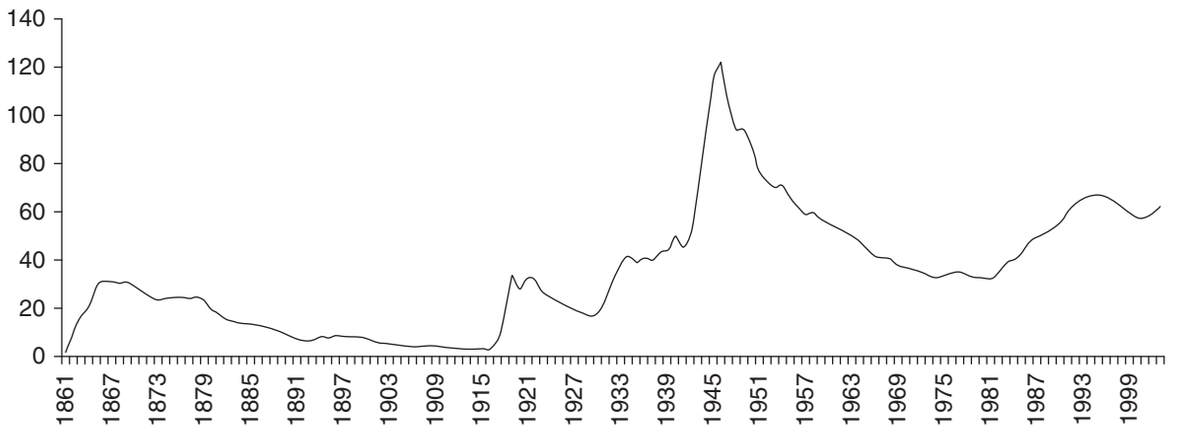


**Figure 36.1** Ratio of U.S. Federal Gross Debt to U.S. GDP, 1861–2003

SOURCE: Noll (2008). Derived from figures provided by Robert Sahr, Oregon State University.

NOTES: The figure shows what the U.S. Treasury and this chapter refer to as Gross Debt. The U.S. Treasury defines the *publicly held debt* to be the Gross Debt minus U.S. trust funds.

large fiscal imbalances that imply large future tax liabilities (more on this point subsequently).

Fourth, state and local governments include bond finance of capital infrastructure expenditures in capital budgets. In contrast, the U.S. government does not report a capital budget. Instead, it includes debt finance of capital expenditure in the Gross Debt. However, to the extent that public capital is productive, it provides taxpayer benefits that tend to offset the liabilities. Productive public capital can increase national income, generating tax revenue that reduces future borrowing needs. The 2008 *Economic Report of the President* reports that federal government net investment was $38 billion in 2007.

Fifth, most outstanding U.S. debt held by the public has a fixed dollar value. However, inflation reduces the real value of this debt. This means that inflation reduces the real value of tax revenue that will be needed in the future to finance debt service. Thus, the current dollar value of the debt overstates its real economic effect. This should not be taken to imply that the national government has license to reduce its real debt by expansionary money policy that would create inflation. For, taxpayers learn to expect persistent inflation. Expected inflation causes nominal interest rates to increase, which causes the dollar value of the debt to rise. Everything else constant, over time, inflation loses its ability to decrease the real value of debt.

Sixth, Laurence Kotlikoff (1992) argues that proper accounting reveals that the true value of the U.S. government's liabilities far exceeds the Gross Debt. In addition to the government's explicit liabilities counted in the Gross Debt, there are large implicit liabilities created by the Medicare system and Social Security. Future revenues expected to be generated by these programs fall far short of the implicit liabilities. Estimates of the shortfalls vary between $25 trillion (Rosen & Gayer, 2008) and $40 trillion (Kotlikoff & Burns, 2004). However, in the future, the U.S. government could legislate lower benefits, whereas it almost certainly will not default on its bonds. Thus, the true size of the U.S. Gross Debt remains a matter of debate.

## What Is a Budget "Deficit" and What Is Its Relation to the Debt?

A national government budget deficit occurs when expenditure exceeds tax and fee revenue in a single fiscal year. In this case, the government must borrow to make up the difference. Therefore, a deficit causes the Gross Debt to increase. For example, U.S. federal government expenditures exceeded its revenues by $363 billion in 2007. Thus, the Gross Debt increased from about $8.5 trillion at the end of 2006 to about $8.9 trillion at the end of 2007.

The official deficit reported by the U.S. Treasury in 2007 was $162 billion. Note that this is less than the $363 billion figure just cited. The smaller figure results from the fact that the official deficit includes the "surplus" in the Social Security program. The government spends Social Security surpluses each year. Because the surpluses are spent, when Social Security receipts eventually fall short of benefits paid (around 2017), the difference must be made up by public bond sales, or by increased taxes, or by decreased spending. Public finance economists argue, therefore, that a more accurate accounting of growth in the debt would exclude Social Security surpluses from the deficit. In this case, the $363 billion figure is a more precise measure of the increase in the debt than the official deficit reported by the Treasury.

Recall that the absolute dollar amount of the national debt does not convey its impact on the economy because the size of the economy grows over time. Economic scale also affects the economic impact of a country's deficit. The preceding figure indicates that the Gross Debt/GDP ratio reached its highest recorded value at the end of WWII. At that time, the deficit/GDP ratio also reached its highest value, about 13%. In 1983, the ratio reached a postwar record of 4.9% (see Figure 36.1). In 2008 the deficit/GDP had fallen to about 2.8%. However, at this writing, the Congressional Budget Office has predicted that the recession that began at the end of 2007 could drive the ratio above 8% in fiscal 2009.

## Do Budget Deficits Improve Economic Welfare?

Almost without exception, U.S. federal government deficits have been a fact of life since the early 1960s. Partly as a result of this record, a vigorous debate over the costs and benefits of deficits has arisen. Deficits have potential to both improve equity and diminish it. They can create efficiencies as well as inefficiencies. This section reviews the arguments pro and con.

It is useful to begin with what the debate is *not* about. Deficit expenditures on productive public capital investment can be efficient in cases where private markets underproduce such investment. For the most part, the deficit debate is not about debt finance of public capital infrastructure. Instead, the discussion in this section focuses on deficit finance of the national government's operating expenditures.

The economics of budget deficits are somewhat different in wartime and peacetime. Consider wartime. Wars tend to absorb outsized amounts of an economy's resources. Rather than relying solely on taxes to acquire resources, national governments often resort to borrowing (see Figure 36.1). The alternative is to increase taxes. Because the wartime generation sacrifices the use of resources, while future generations presumably benefit from the wartime effort, it has been argued that borrowing to finance wars is equitable. By the same token, it has been argued that debt finance of a war is an efficient investment if it makes citizens in current and future generations better off than they otherwise would be.

The economics of peacetime deficits are more complicated. First, the effects of deficits appear to depend on the specific types of expenditures the deficits support. Second, the effects of deficits appear to depend on the

way household saving responds to deficits. Third, the impact deficits have depends on where the economy is located in the business cycle when the deficit is incurred.

Before discussing these issues, it is useful to address a general problem that may arise in the case of debt finance of most types of public expenditures. If current taxpayers believe that others, not they, must repay funds borrowed by the government, debt finance tends to be more politically palatable than taxes. In this case, the ability to finance debt may lead to larger government expenditures than taxpayers would choose if they believed, instead, that they themselves would be required to repay the debt. Therefore, the government's ability to use deficit finance can lead to inefficiently large amounts of expenditure.

This inefficiency may be offset, to some degree, by the fact that debt finance can be used to reduce the variability of tax liabilities. Reduced tax variability is more efficient (Barro, 1979). To see how, note that variation in tax liabilities from period to period reduces taxpayer welfare, even if the overall tax liability is unchanged. This is true because variability in tax liability causes variability in disposable income, which, in turn, causes variability in household consumption. By the well-known economic law of diminishing returns, an increase in consumption adds less to welfare than a decline in consumption reduces welfare. Consider two ways a household can allocate a given level of income: In the first case, consumption is high at first, followed by low consumption; in the second case, the same level of income is consumed evenly over time. The second case provides more utility. It follows that a given level of tax liabilities reduces utility by less, if tax liabilities are constant over time. In the short run, government revenue requirements vary greatly because of short-run variation in demand for government services. Altering tax liabilities each period to equal the changing revenue requirements reduces welfare more than smooth tax liabilities. Debt finance can be efficient if it is used to smooth tax liabilities: The government can maintain constant tax liabilities by borrowing when revenue requirements are relatively high (e.g., in wartime or in economic crisis) and by repaying loans when requirements are relatively low.

Kotlikoff (1992) argues that the U.S. pay-as-you-go Social Security system incurs large *implicit* deficits. If the Social Security program is not reformed, current generations are receiving Social Security benefits that must be paid for by future generations, transferring wealth from future generations to current generations. In this case, implicit deficit finance is inefficient: Resources are not allocated where they have the highest social value in the sense that if all citizens, future and current, were fully informed about the system's costs and benefits, they would chose a different allocation. The Social Security system also is inequitable because citizens who are born into different generations, but have equal abilities to pay, face different levels of taxation.

Further, deficit finance, implicit as well as explicit, has a potential to inefficiently reduce future production and a nation's standard of living (Altig, Auerbach, Kotlikoff, Smetters, & Walliser, 2001; Auerbach & Kotlikoff, 1987; Feldstein, 1974). This point requires some explanation.

Deficits can reduce the future standard of living if they increase consumption and reduce national saving. To see this, note that national saving is the sum of private, business, and government saving. Also note that the economic effect of borrowing is equivalent to a decline in saving. For example, a household can finance a car by reducing its bank account by $25,000, reducing savings, or by retaining the bank account and borrowing $25,000: The economic effects are the same. Therefore, a deficit is a decline in government saving. If a deficit-induced decline in government saving is not accompanied by an equal increase in private saving, national saving will be lower than otherwise. Lower national saving tends to increase the interest rate: In this case, investment declines because the interest rate is a principal part of the cost of capital. The process by which increased deficits cause investment to decline is called *crowding out.*

If deficits crowd out private investment, and the government does not use the borrowed funds to purchase productive capital, national capital investment will be lower than it otherwise would be. In this way, crowding out can reduce the amount of capital inherited by future generations, thereby lowering future standards of living. Douglas Elmendorf and N. Gregory Mankiw (1999) estimate that deficits have reduced U.S. incomes by 3% to 6% annually. In this case, unless there is an offsetting economic rationale to run deficits, they are inefficient.

However, deficits may not have such dire effects and, Robert Barro (1974) argues, may not matter at all. Barro argues that deficits (implicit or explicit) do not affect national saving, capital investment, or the future standard of living. This will be true if private savers respond to the deficits by saving more, thereby offsetting the deficit's negative effect on national saving. Private savers might increase saving if they foresee the negative effect the deficit otherwise could have on future living standards, and if they are determined to prevent deficits from undermining their children's futures. In opposition to this idea, B. Douglas Bernheim (1987) argues that deficits reduce national saving if the amount that households can borrow is restricted by financial markets. In this case, deficit-financed tax cuts provide these households with funds they would like to borrow from markets, but cannot. These households treat deficit-financed tax cuts as they would borrowed funds, and spend them, reducing private saving and national saving.

Whether debt finance does or does not reduce national saving is an empirical issue and can be decided only on the basis of careful statistical analysis of savers' behavior (Bernheim, 1987). Currently, there appears to be fairly widespread agreement among economists that deficit finance tends to reduce national saving. In particular, there appears to be a consensus that deficits increase aggregate demand. Economists who take this view argue that if

deficit finance is used to increase government expenditures, aggregate demand increases directly. If deficit finance is used to reduce taxes, taxpayers spend at least part of the tax cut, increasing aggregate demand indirectly. The remainder of the discussion studies the effects of deficits in these cases.

Tax collections based on personal income, corporate income, and sales decline during contractions. At the same time, social spending by government tends to rise during contractions. Thus, if the federal government were prohibited from running deficits in contractions, it would be required to increase taxes or cut spending, or both. The federal government's budget is a very large share of the U.S. economy (about 20% of GDP), so such tax increases or spending decreases would tend to exacerbate contractions, which would inefficiently increase the extent of unemployed resources. Thus, contraction-induced deficits are favored by many macroeconomists and policy makers.

Are contraction-induced deficits consistent with the government's IBC? As explained before, the important economic implication of IBC is that government must reliably pay the market rate of return on its debt; thus, the debt cannot grow faster than the economy in the long run. If the government ran persistent deficits, the debt would outrun the economy's ability to produce tax revenue required to service the debt, so persistent deficits are not sustainable. Thus, contractionary deficits satisfy IBC only if the government eventually runs surpluses. An efficient deficit policy would have the government repay recessionary deficits with expansionary surpluses. In this case, the government budget would be balanced *across the business cycle:* Resources absorbed by the government during recessionary deficits would be freed up during subsequent expansions.

Because resources are underutilized during recessions, many macroeconomists argue that the government can and should be more proactive than simply permitting recessionary deficits. They argue that the government should purposely increase deficits during economic contractions, and that doing so would increase aggregate demand, thereby reducing the severity of contractions. In this case, policy makers would resist cyclical contractions by legislating larger expenditures and lower taxes. This is called *activist counter-cyclical* fiscal policy. Such policy has a potential to be efficient, if it does not short-circuit the economy's self-correcting mechanisms and if the government balances its budget over the business cycle.

## How Are Government Debt and Foreign Debt Related?

The *foreign debt* is the accumulated amount that foreigners have lent to both private and government borrowers in the United States, minus the accumulated amount foreigners have borrowed from (primarily private) U.S. lenders. A substantial fraction of loans to U.S. governments are provided by foreigners. For example, before the 2008 financial crisis set in, about 32% of U.S. Gross Debt was owned by foreigners.

As explained earlier, budget deficits tend to increase interest rates, particularly if the economy is expanding. The increase in interest rates encourages foreign lending to the United States, which increases the U.S. foreign debt. Thus, U.S. government debt tends to be positively correlated with the U.S. foreign debt.

## Does Foreign Ownership of U.S. Debt Affect the Standard of Living?

By themselves, deficits tend to increase interest rates. Foreign lending increases the supply of loanable funds, which tends to offset the positive pressure deficits exert on interest rates. More generally, foreign lending tends to reduce the cost of U.S. borrowing, which can support economic growth. However, foreign lending also tends to increase the foreign exchange value of the U.S. dollar because lenders must buy dollars in order to buy the U.S. debt. An increase in the exchange value of the dollar tends to reduce U.S. exports and increase U.S. imports, which tends to reduce economic growth. Therefore, the net effect that foreign lending has on the U.S. standard of living is an empirical question and cannot be determined by theory alone.

Many economists assume that foreign lending increases growth and the standard of living, at least in the short run. If so, the effect may be temporary: After all, interest on foreign debt must be paid, and foreigner lenders may someday choose not to refinance loans to the U.S. Does government borrowing from foreign investors increase the U.S. standard of living in the long run? There are a number of possibilities.

First, if U.S. governments use the borrowed funds for productive capital investment, the capital stock will be larger and future U.S. production will be larger. If the productivity of the public capital investments exceeds the interest on the loans, future U.S. living standards will tend to be higher than otherwise. However, if interest payments exceed the productivity of public capital, future U.S. living standards would tend to be lower than otherwise, even if U.S. production is larger. In this last case, even though the borrowed funds increase U.S. output, the income generated contributes to foreign living standards, not U.S. living standards.

Second, if U.S. governments use the borrowed funds for consumption, the foreign debt does not contribute to future production, and debt service would reduce future U.S. living standards because part of the unchanged level of production must be delivered to foreign lenders.

Third, a large and rapid shift away from foreign lending to the U.S. could cause the foreign exchange value of the dollar to decline precipitously. In this case, U.S. interest rates would rise, and the U.S. standard of living could decline sharply. Fortunately, events such as these have never occurred in the U.S. However, they are a real risk

whose possibility should be taken seriously in fiscal policy makers' deliberations.

### A Balanced Budget Amendment?

The federal government is not required to balance its budget. Because deficits can be used to increase government expenditure above the amount fully informed taxpayers would choose, some public policy analysts advocate a strict Balanced Budget Amendment (BBA) to rule out federal deficits.

A disadvantage of a strict BBA is that it would prohibit deficit spending in recessions. This would require the government to increase taxes or cut spending during economic contractions, which would tend to exacerbate downturns and make taxes more variable, both of which are inefficient. Further, a strict BBA would prohibit countercyclical fiscal policies of the sort that observers of many political persuasions support during recessions.

### Are U.S. Deficit (Debt) Policies Sustainable?

The crucial economic implication of the IBC is that government debt cannot persistently grow faster than the economy. If the debt growth persistently exceeds GDP growth, investors would eventually demand prohibitively high interest rates, or simply refuse to purchase additional government debt. The relative growth rates of debt and GDP are reflected in the debt/GDP ratio, which increases if the debt grows faster than the economy. In 1980, 1990, 2000, and 2008, the debt/GDP ratio was, respectively, 32.0%, 55.2%, 57.3%, and 73.0%. For 3 decades now, U.S. debt policy has been on an unsustainable path. The longer the U.S. continues on this trend, the more serious the economic repercussions will be.

## Conclusion

Government budgets provide information the public needs to evaluate costs and benefits of government services and indicate whether the government is operating in a financially responsible manner. Capital budgets allow taxpayers and investors to evaluate the risk and returns of government investment more easily. State and local governments report capital budgets. The U.S. federal government does not. A federal capital budget would make it easier for taxpayers and investors to evaluate the risk and net returns of federal government programs.

State and local governments can reduce the adverse budget effects of economic contractions by properly structured and appropriately sized RDFs. RDFs can help reduce the need for ad hoc reductions in expenditures and increases in taxes that often accompany budget shortfalls. Thus, RDFs can reduce the risk of credit downgrades, uncertainty, and the excess burden of state and local taxes.

The National or Gross Debt is the value of all outstanding national government bonds. If government debt persistently grows faster than the economy, creditors will demand abnormally high interest rates or may refuse to purchase government debt. The economy can suffer severe damage. In a debt crisis, the government must raise taxes or cut government expenditure, default on its liabilities, or repudiate the debt.

There is a long list of reasons why the reported value of national debt must be interpreted cautiously, if one is to understand its economic effects. First, the debt/GDP ratio adjusts the raw value of the debt for economic scale and provides a reasonable tool that can be used to compare debt policies in different countries, as well as the debt of a single country in different years. Second, central banks hold substantial fractions of national debt: This tends to diminish some of the adverse effects of government debt and should be taken into account in attempts to evaluate debt's economic impact. Third, some fraction of debt finances public capital, which can contribute to economic growth and tax revenue needed to service the debt. Thus, government infrastructure investment also should be accounted for when evaluating the effects of debt. For this reason, many economists favor a federal government capital budget separate from the operating budget. Fourth, a substantial part of the debt has a fixed dollar value, which declines in real terms when inflation is positive. Thus, many economists agree that the real value of the debt influences interest rates, investment, and the standard of living. Fifth, the U.S. government has incurred large implicit liabilities that are not counted in the Gross Debt. Thus, the reported national debt understates the government's true financial liabilities and provides a misleading estimate of future tax revenues necessary to sustain U.S. fiscal policy.

Debt issue can be used to reduce the variability of taxes, increasing efficiency. It can be used to spread out the costs of government to future generations, who benefit from current government expenditures, which most observers judge to be equitable. However, the ability to issue debt can lead to inefficiently large government expenditure if voters believe they will benefit from spending they will not pay for. Implicit government borrowing can also be used to transfer wealth from future to current generations, which can be inefficient and inequitable (Kotlikoff, 1992). Debt finance can reduce future production and a nation's standard of living, if deficits crowd out private investment and government does not use the borrowed funds to purchase public capital.

Many macroeconomists argue that the government can and should use countercyclical deficit policy to reduce the severity of economic contractions. Countercyclical fiscal policy has a potential to be efficient, if it does not short-circuit the economy's self-correcting mechanisms, and assuming the government balances the budget over the business cycle.

Deficit-induced increases in interest rates encourage foreign lending to the United States, which increases the foreign debt. Foreign lending tends to reduce borrowing costs and can contribute to economic growth. However, foreign lending also tends to increase the foreign exchange value of the U.S. dollar, which can reduce U.S. net exports, reducing growth somewhat. The net effect cannot be determined by theory. Funds borrowed from foreign investors and used for public investments can increase the standard of living in the long run, if the return on investment exceeds the interest cost. If the funds are used for consumption, foreign lending will tend to reduce future U.S. standards of living. Foreign debt runs a risk that foreigners may suddenly shift investments away from the United States, which can damage the U.S. economy severely.

Finally, government debt cannot grow faster than the economy in the long run. U.S. Gross Debt has grown faster than the economy for decades. U.S. fiscal policy surely is not sustainable.

## References and Further Readings


Altig, D., Auerbach, A. J., Kotlikoff, L. J., Smetters, K. A., & Walliser, J. (2001). Simulating fundamental tax reform in the United States. *American Economic Review, 91*(3), 575–595.

Auerbach, A. J., & Kotlikoff, L. J. (1987). *Dynamic fiscal policy.* Cambridge, UK: Cambridge University Press.

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.

Bassetto, M., & Sargent, T. (2006). Politics and efficiency of separating capital and ordinary government budgets. *Quarterly Journal of Economics, 121*(4), 1167–1210.

Bernheim, B. D. (1987). Ricardian equivalence: An evaluation of theory and evidence. *NBER Macroeconomics Annual, 2,* 263–304.

Blocker, A. W., Kotlikoff, L. J., & Ross, S. A. (2008). *The true cost of Social Security.* Unpublished working paper, Boston University Economics Department.

CIA World Factbook. (2008, January). Available from https://www.cia.gov/library/publications/the-world-factbook/index.html

*Economic report of the president. Transmitted to the Congress February 2008, together with the annual report of the Council of Economic Advisers.* (2008). Washington, DC: Government Printing Office. Available from http://www.gpoaccess.gov/eop/2008/2008_erp.pdf

Elmendorf, D. W., & Mankiw, N. G. (1999). Government debt. In J. B. Taylor & M. Woodford (Eds.), *Handbook of macroeconomics* (pp. 1616–1669). Amsterdam: Elsevier North-Holland.

Feldstein, M. (1974). Social Security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy, 82,* 905–926.

Fisher, R. C. (2007). *State and local public finance.* Mason, OH: Thomson South-Western.

Fox, W. F. (1998). Can the state sales tax survive a future like its past? In D. Brunori (Ed.), *The future of state taxation* (pp. 33–48). Washington, DC: Urban Institute Press.

Gokhale, J., Page, B., Potter, J., & Sturrock, J. (2000). Generational accounts for the U.S.: An update. *American Economic Review, 90*(2), 293–296.

Kotlikoff, L. J. (1992). *Generational accounting: Knowing who pays, and when, for what we spend.* New York: Free Press.

Kotlikoff, L. J., & Burns, S. (2004). *The coming generational storm.* Cambridge: MIT Press.

Lindsey, B. L. (1987). Individual taxpayer response to tax cuts: 1982–84: With implications for the revenue maximizing tax rate. *Journal of Public Economics, 33,* 173–206.

Mankiw, N. G. (2007). Government debt. In *Macroeconomics* (pp. 431–452). New York: Worth.

Noll, F. (2008, August). The United States public debt, 1861 to 1975. In R. Whaples (Ed.), *EH.Net encyclopedia*. Available from http://eh.net/encyclopedia/article/noll.publicdebt

Rosen, H. S., & Gayer, T. (2008). *Public finance.* New York: McGraw-Hill Irwin.

Russo, B. (2002). *Report on North Carolina State revenue shortfalls and budget stabilization funds.* Unpublished manuscript, University of North Carolina, Charlotte, Economics Department. Available from http://belkcollegeofbusiness.uncc.edu/brusso

# 37

# Monetary Policy and Inflation Targeting

Pavel S. Kapinos
*Carleton College*

David Wiczer
*University of Minnesota*

Inflation targeting (IT) is a framework for the conduct of monetary policy, under which the monetary authority announces a medium- or long-run inflation target and then uses all available information to set its policy instrument, the short-term nominal interest rate, so that this target is met. Short-lived deviations from the inflationary target may be acceptable, especially when there may be a short-run trade-off between meeting the target and another welfare consideration, for example, the output gap—the difference between actual and potential output. Hence, although the central bank commits to meeting a certain inflationary target, in practice, IT takes a less rigid form, with the central bank exercising some discretion over the path of actual inflation toward its target. Recently, dozens of central banks around the world have introduced IT as their operational paradigm. Numerous studies indicate that this policy has been successful in achieving macroeconomic stability at no long-run cost in terms of lower real activity. Many central banks that have not explicitly subscribed to IT have been shown to follow it implicitly.

IT provides a way for the central bank to communicate its intentions to the public in a clear, unequivocal manner, making the conduct of monetary policy more transparent and predictable. Transparency allows the public to hold the central bank accountable for its policy actions. In fact, in some countries, inflation-targeting central banks are subject to intense public scrutiny from the legislative bodies. Predictability of monetary policy allows the central bank to manage public inflationary expectations and better anchor them around the inflationary target; this allows the central bank to achieve macroeconomic stability more effectively.

Alan Greenspan (2004) has famously posited that "the success of monetary policy depends importantly on the quality of forecasts" (p. 39). In practice, monetary policy affects macroeconomic activity with a lag, hence forming accurate forecasts of the future macroeconomic activity is of paramount importance. Several IT central banks are reforming their communications procedures to make their macroeconomic forecasts, as well as projections of the nominal interest rate path to meet their stated objectives, publicly available.

Although, historically, central banks have paid close attention to a large number of macroeconomic variables, there are several reasons for the central bank to focus on only one variable as its target and to make inflation that variable. The central bank generally has only one independent instrument: the short-term nominal interest rate. With this single instrument, the Tinbergen principle states that they can achieve their target for only a single variable. Even though they may rightly care about many macroeconomic variables, they would need new tools to move each one independently. Furthermore, theoretical literature

focusing on the role of forward-looking (rational) inflationary expectations formed by the public has shown that central banks that take a particularly tough stance against inflation may generate better macroeconomic performance in the long run (Rogoff, 1985). Inflationary conservatism on behalf of the central bank moderates inflationary expectations and brings about full employment at a lower rate of inflation.

Beyond its theoretical advantages, IT succeeds in practice with lower, more stable inflation and more stable real activity. As more countries adopt the IT framework, the evidence in its support grows. After New Zealand in 1990, a cascade of developed and developing countries followed, including relatively large economies, such as Canada and Brazil, and relatively small ones, such as the Czech Republic and Israel. Although the largest central banks, the European Central Bank (ECB) and the U.S. Federal Reserve, have yet to adopt IT explicitly, in practice, their monetary policy conduct seems to match the IT framework quite closely, with Ben Bernanke, the Chairman of the Federal Reserve, known as an early vocal proponent of explicit IT. Its growing popularity clearly signals the merits of adopting IT as the paradigm for the conduct of monetary policy.

This chapter's assessment of modern IT first traces its historical roots, exploring both the antecedent monetary frameworks and the developments in economic theory that motivated its present-day form. We then discuss how IT is implemented in practice and the preliminary empirical evidence from the countries that have adopted it. To conclude, we offer suggestions on how the IT consensus might grow and its prospects as a unifying framework for monetary policy.

## Historical Background

In 1931, faced with the global depression and speculative attacks on its currency's gold peg, the Swedish Riksbank began a 6-year experiment with price level targeting (PLT), the first and only of its kind. Just as the gold standard was renounced, the bank promised price stabilization "using all means available." Like modern IT, PLT sought to commit monetary policy to a single nominal goal. The Riksbank was a likely pioneer, given the country's economics tradition. As early as 1898, Knut Wicksell, a Swedish economist, discussing the role of the monetary instrument, recommended that the discount rate should move with the price level to dampen its volatility. When its PLT policy was first formulated, the Riksbank was given instrument independence—that is, it could use interest rates any way it saw fit to meet the price level target set by the government.

The price-targeting regime was, however, conceived in a very different context than that faced by the modern IT adopters. In the depressed economic conditions of the 1930s, deflation was a serious risk, and even though many feared inflation after Sweden eschewed the gold standard, adhering to the price level target was also meant to fight deflation that crippled the rest of the world. Indeed, from 1928 to 1932, Swedish prices declined steadily in line with the rest of the world but then rose slowly for the rest of the decade, and though unemployment peaked in 1933, output stayed comparatively strong through the 1930s.

As a proto-targeting regime, the Swedish experiment differed in several ways from modern IT. By committing to hold a specific price level, excess inflation has to be compensated with periods of deflation. In contrast, modern IT does not compensate for inflationary buildup in prices, as long as the rate of inflation returns to the target. The Swedish experiment was initially meant to be only a temporary break from the gold standard. They announced the policy with the caveat that it was a temporary measure, but without a strict ending date. Modern inflationary targeting, on the other hand, can be successful only in the long term, because much of its performance depends on how well the central bank can build credibility and manage the public's inflationary expectations. To further muddle policy design, the parliament gave the Riksbank its targeting mandate but did not communicate a quantitative target, nor even a price measure (i.e., CPI, wholesale prices, etc.). And while an IT central bank today uses forecasts to communicate its intentions, the Riksbank never published these. The Riksbank's PLT experiment ended in 1937 when monetary and fiscal policies were reorganized along Keynesian lines to pursue aggregate demand stabilization.

Keynesian thinking dominated policy right through the 1960s, and policy prioritized full employment above price stability. Fitting its diminished importance globally, monetary policy was squeezed into a tight framework after World War II. World prices were anchored to the dollar, which hypothetically was convertible to gold. Central banks still worried about inflation but believed that it could be easily traded off against unemployment in a fashion that did not change over time. A coordinated action of taxing, spending, and discount rate manipulation would pinpoint a position on the so-called Phillips curve, an empirical regularity characterizing the trade-off between inflation and unemployment that policy makers felt could be exploited to match social preferences with respect to these two variables.

The major question became gauging what combination of evils, unemployment versus inflation, would be most socially tolerable. Not surprisingly, the generation that still remembered the Great Depression was more sensitive to unemployment. The popular press was sure of the economists' prowess, proclaiming in a 1965 *Time* cover story that successful policy could coax an economy to full employment. The story was titled "We Are All Keynesians Now," a quote from Milton Friedman[1], but it never mentioned policy's role in guiding inflationary expectations. In the 1950s and 1960s, macroeconomists rarely objected to linking unemployment and inflation with some linear

relation and estimating the equation's coefficients. This led to a natural policy prescription that anytime lower unemployment is desired, inflation should be allowed to increase. The working assumption was that inflation would move along a static Phillips curve, given the rate of unemployment.

An undercurrent of economists in the 1960s and a deluge in the 1970s called to question the status quo that had developed up until then. Accelerating inflation in the United States spread across the world, and the economics profession began to question whether these reduced-form equations really were as time invariant as Keynesians claimed. They tried to ground macroeconomic relationships in models of rational individuals and found Keynesian economics critically incomplete and misleading. The result was a Phillips curve incorporating inflationary expectations that became central in explaining the dynamic macroeconomic evolution.

In his celebrated presidential address to the American Economic Association, Friedman (1968) suggested that a monetary policy that actively tries to promote macroeconomic stability may actually disrupt it because of the poor quality of information available to policy makers. He suggested targeting monetary aggregates (by keeping money supply growth constant) to promote long-term stability. Friedman's speech almost perfectly coincided with the breakdown of the Bretton-Woods system in the early 1970s, in whose wake central banks were forced to design new operational paradigms for the conduct of monetary policy. This sort of policy is an "intermediate target" because money supply is not a welfare criterion in itself but is a major determinant of inflation.

Money targeting had to be abandoned relatively quickly due to the unstable relationship between monetary and other macroeconomic aggregates. There is widespread consensus that the Federal Reserve pursued monetary targets only between 1979 and 1982. Instead, central banks have started searching for alternative frameworks for the conduct of monetary policy, and toward the 1990s, a number of central banks were poised to start IT.[2] Before we document the adoption of IT by central banks around the world, we will lay out a simplified version of the theoretical framework that is frequently used for analyzing the design of an IT regime.

## Implementing IT: Theory

A large body of recent theoretical literature has been dedicated to the study of monetary policy. Richard Clarida, Jordi Gali, and Mark Gertler (1999) and Michael Woodford (2003) provide a comprehensive overview of the issues and methods involved in studying monetary policy conduct from the New Keynesian perspective. The baseline version of the model has three sectors: households, firms, and the monetary authority. The first-order conditions from the household utility maximization problem give rise to the IS schedule that describes the equilibrium evolution of a measure of real activity or *output gap,* defined as the percentage deviation of actual output from potential, as a negative function of the real interest rate. By adjusting the nominal interest rate, therefore, the central bank can influence the level of real activity in the economy.

Monopolistically competitive firms set optimal prices for their output, with different firms resetting prices at different points in time, which gives rise to changes in relative prices of different outputs and hence welfare-loss-generating distortions that the policy maker is called to moderate.[3] The first-order condition to this problem yields the New Keynesian Phillips curve (or the aggregate supply relation); a key feature of this structural equation is the presence of forward-looking, rational inflationary expectations. The baseline version of this Phillips curve takes the form

$$\pi_t = \beta \pi_{t+1|t} + \kappa (y_t - y_t^n) + \varepsilon_t,$$

where $\pi_t$ is inflation, $\pi_{t+1|t}$ is the expectation of next period's inflation based on information available at the present time period, $(y_t - y_t^n)$ is the (log) difference between actual and natural (or potential) output—or output gap—and $\varepsilon_t$ is the cost-push shock. The *output gap* term can be motivated by changes in the firm's labor costs, whereas the cost-push shock is by changes in the markup that firms charge in excess of their labor costs. The presence of the *expected inflation* term is due to price stickiness modeled by the firm's potential inability to change prices optimally in the future. Anticipating future inflationary pressures then enters the firms' current pricing decisions and makes it important for the central bank to engage in managing inflationary expectations, as moderating those will result in lower inflation in the present.

Finally, the central bank needs to determine how the nominal interest rate is set in order to close this three-sector model. There are two ways of thinking about the conduct of monetary policy that have been widely used in the recent literature. One is due to the insight of John Taylor (1993), who pointed out that the Federal Reserve seemed to set the nominal interest in response to inflation and output gap. A number of empirical studies have confirmed that following this so-called Taylor rule has characterized the conduct of monetary policy in the United States in the recent past quite well. A large body of theoretical literature has shown that Taylor rules deliver macroeconomic stability in a variety of models.

Another way that monetary policy has been modeled is by assuming that the central bank minimizes a discounted stream of weighted averages of squared deviations of inflation from its target level and output from its target level (set at the potential level of output). Lars Svensson (2002, 2003) has been a particularly strong advocate of thinking about monetary policy in terms of minimizing volatility of inflation and output around their target levels rather than

the mechanistic approach exemplified by Taylor rules. (Svensson calls the former *targeting rules,* because they are built around inflation and output targets, and the latter *instrument rules,* because they explicitly define the value of the nominal interest rate in terms of the other variables.) This setup places the conduct of monetary policy into a dynamic setting and offers a richer variety of insights.

One such key insight is the problem of the time inconsistency of optimal policy, originally pioneered by Finn Kydland and Edward Prescott (1977) and Guillermo Calvo (1978). Suppose that the economy experiences a positive transitory cost-push shock. The central bank can moderate its instant inflationary effect by promising to keep output below potential in the future time period. This reduces inflationary expectations and (partially) offsets the effect of the cost-push shock. Hence in the present, this shock will generate lower inflation and a smaller decrease in output below potential. In a way, the central bank commits to spread the real-economic pain from the cost-push shock over several time periods, even when the immediate impact of the shock is long gone. Although this policy generates optimal outcomes over the long haul, it does run into short-run difficulties. In particular, in the time period following the transitory shock, the central bank has no incentive to make good on its promise to generate a recessionary environment, which helped moderate inflationary expectations—that is, it has an immediate incentive to renege on its contractionary promise because the effects of the shock have already dissipated. But if it discretionarily reneges on its promise, once another cost-push shock materializes, it will not be able to pursue a dynamically optimal stabilization policy with any measure of credibility, hence generating worse macroeconomic outcomes from there onward.

Another implication that emerges out of the dynamic targeting setup is that, especially given the long and variable lags in the transmission of monetary policy through the economy, it makes little sense for a central bank to set the nominal interest rate in response to a small number of contemporaneous variables in a predetermined, mechanical fashion. Furthermore, given the forward-looking nature of macroeconomic agents, what may matter more for the successful conduct of monetary policy is not the current level of the nominal interest rate but its projected path over some future time horizon. Identifying such a path is impossible without first postulating a dynamic objective over which the central bank wants to optimize.

In practice, these challenges make it virtually impossible for a central bank to achieve its target with current inflation, and instead, it will target expected inflation some periods into the future. Targeting future inflation also gives the central bank greater flexibility in considering output volatility as well. Woodford (2003) explains,

> The argument that is typically made for the desirability of a target criterion . . . with a horizon $k$ some years in the future is that it would be undesirable not to allow temporary fluctuations in inflation in response to real disturbances, while the central bank should nonetheless provide clear assurances that inflation will eventually be returned to its long-run target level. (p. 292)

IT central banks seem to exhibit their most meaningful differences by the length of this lag. More "flexible" IT regimes accommodate some real shocks and often will not return inflation to its target for much longer than their "stricter" counterparts who target the earliest feasible date. Svensson (1998) estimates that this time range for meeting the target is anywhere from 1.5 to 2.5 years. (As discussed below this "flexibility" opens these central banks to the criticism that their policy making is actually arbitrary.) More flexible IT regimes may allow the monetary authority to give some attention to unemployment or a real objective while still promising a certain level of inflation in the long run. Additionally, they may also allow greater instrument stability, as wide swings in interest rates may be painful in themselves. To show how the central bank balances these additional criteria, the terms that characterize their stabilization can be added into the central bank's objective function with weights to show their relative importance. In a more strict IT regime, other concerns have relatively low weights, so that it may worry about large swings in these variables, but it still chiefly focuses on inflation stabilization.

Some critics of IT insist that insofar as the duration of actual inflation being off its target is uncertain, IT affords the monetary authority too much flexibility and is therefore destabilizing. For instance, in his criticism, Benjamin Friedman (2004) invokes the Tinbergen principle, which holds that with only one independent instrument, the variables relevant to monetary policy can be expressed in relation to one particular variable. This means that when a shock hits, and output and inflation are both off their desired level, the transition path of output is dependent on the path of inflation, so by choosing how long inflation may be out of the target range, the central bank implicitly chooses how output will respond. Since IT states the target for only one variable—inflation—and fails to communicate the central bank's preferences with respect to other variables of interest, it may be viewed as too opaque. This critique, however, does not breach the consensus view that IT is successful in augmenting the central banks' transparency, accountability, and success in mitigating business cycle volatility.

Although a tremendous amount of progress has been made in identifying the chief challenges of monetary policy making and demonstrating the benefits of IT, there does not exist a universal consensus on the exact methods for the implementation of IT. The next section reviews the recent experience of different countries whose central banks have adopted IT. Despite considerable differences in implementing this policy, the picture that emerges

from our description points to its success in achieving macroeconomic stability.

## Implementing IT: Practice

New Zealand was first to set an explicit inflation target for the medium run. The Reserve Bank of New Zealand (RBNZ) had a wide-ranging mandate from a 1964 law that charged it with "promoting the highest degree of production, trade and employment and of maintaining a stable internal price level." Inflation stayed in double digits for most of the 1970s and 1980s, a cumulative 480% between 1974 and 1988. But by the time IT was introduced in 1990, a period of very high interest rates had inflation mostly under control. Ben Bernanke, Thomas Laubach, Frederic Mishkin, and Adam Posen (1999) believe that this timing was important to build confidence in the IT framework, as the public did not blame the new policy regime for the induced tight monetary conditions.

The inflation target was in fact a range set by the government that gradually fell from 2.5% to 4.5%, to 1.5% to 3.5%, and finally to 0% to 2% after December 1993. Though the government held the central bank accountable to keep inflation in the middle of a range it set, the RBNZ enjoyed instrument independence—that is, the central bank could independently choose its policy for interest rate and exchange rates. The IT framework in New Zealand also evolved to become less strict, in that deviations are now allowed to persist longer than the initial yearly horizon.

IT in New Zealand was also pioneering in the degree of transparency and the central bank's communication of its decision making to the public. The RBNZ is legally required to produce a *Monetary Policy Statement* twice a year that outlines policy decisions. It also publishes forecasts in an *Annual Report,* other research in an annual *Bulletin* and a bi-annual *Financial Stability Report.* Starting in 1996, it has released a *Monetary Conditions Index* to summarize its forecasts.

Despite the eventual success of IT, the RBNZ did initially encounter some difficulties in its implementation. It first tried using "headline" or all-goods consumer price index (CPI) but found this too volatile. The RBNZ developed a concept of *underlying inflation* similar to a core measure to avoid including prices that were themselves sensitive to the first-round effects of a cost-push shock. Unfortunately, the core's components changed occasionally, making it somewhat ad hoc, and its time series was marked by definitional breaks. The relatively tight, certainly by the standards of a small open economy, inflation target range of 0% to 2% imposed additional challenges, as the RBNZ was occasionally forced to produce large swings in the interest rate in order to meet its objective. Several times, inflation breached its ceiling, which was lifted to 3% in 1997. Still, the experiment with IT at the RBNZ has undeniably been a success, as chronic and volatile inflation seems safely in the past.

In 1991, Canada was next in the steady succession of IT adopters, following an announcement by the head of its central bank and the Minister of Finance but without a legislative mandate, unlike in New Zealand. Prior to adopting IT, Canada's experience with inflation was less extreme than New Zealand's, but the 1989 inflation rate of 5.5% was deemed unacceptably high. As the policy was adopted, the disinflation was accomplished by high interest rates and a sharper than expected cyclical downturn; on the upside, however, IT accomplished low inflationary expectations. According to Gordon Thiessen (1999), the Governor of the Bank of Canada from 1994 to 2001, inflationary expectations did not fall because the bank promised it but because low rates of inflation were "realized"; however, the promise helped to entrench these gains:

> It is unlikely that the 1991 announcement of the path for inflation reduction had a significant immediate impact on the expectations of individuals, businesses, or financial market participants. On balance, I think that it is the low realized trend rate of inflation in Canada since 1992 that has been the major factor in shifting expectations of inflation downwards. But the targets have probably played a role in convincing the public and the markets that the Bank would persevere in its commitment to maintain inflation at the low rates that had been achieved.

Canada's IT regime focuses on the medium-term inflationary expectations because they are particularly susceptible to short-run shocks as a small, commodity-exporting economy like New Zealand. Following the RBNZ, the Bank of Canada excludes volatile components from their "core" price index that, in its judgment, are transitory. If these excluded shocks are mean-zero, then headline and core will move together in the medium term. This flexibility allows a central bank to avoid the adverse output implications of contractionary policy and is common among other IT regimes.

Canadian policy makers also emphasized a dialogue with the public, which has become important for the transparency of IT regimes. To help formulate policy response to inflationary expectations, they follow the *Consensus Forecasts* of market economists compiled by the private Conference Board of Canada. They have also geared their own publications to be easily digestible. The Bank of Canada produces the *Monetary Conditions Index,* a summary statistic for price-level determinants, and uses more charts in their *Annual Report* to encourage public confidence in the Bank's ability to meet its IT responsibilities.

Following New Zealand and Canada, the ranks of inflation targeters swelled during the 1990s. The United Kingdom and Sweden experienced sharp downturns and high inflation in the early 1990s and adopted IT regimes shortly thereafter. On the other hand, Australia's situation was similar to New Zealand's—a 1991 recession reduced

inflation before their IT regime was enacted. Thereafter, a slew of developing countries created explicit inflation targets, ranging from traditional inflationary basket cases, such as Israel and Brazil, to postsocialist economies, such as the Czech Republic and Poland. Each has its own idiosyncrasies: Some target *all-items inflation* (e.g., Israel and Spain), while others target an underlying or core measure (e.g., Australia and the Czech Republic). Table 37.1 summarizes the list of inflation targeters and the time of adoption of this policy.

The U.K. experience exemplifies the rationale for IT adoption in many industrialized countries. Following an inflationary bout in the 1980s, the British had stabilized their currency by tying it to the Deutsche Mark. But when large-scale speculation forced the authority to break their peg, the Bank of England experienced a great deal of exchange rate volatility and abruptly lost their intermediate target. To stabilize exchange rates and regain an anchor for price stability, they implemented a band for their medium-term inflation target and long-term goal of 2%. After Britain's success, other developed countries, such as Finland, Spain, and Sweden, also adopted inflation targets in response to foreign-exchange pressure.

Transitional and middle-income countries often have more volatile macroeconomic conditions and so may have even more to gain from the stability that IT imparts. Brazil's inflation has fluctuated wildly through its diverse monetary regimes. When its fixed exchange rate regime failed in 1999, it began IT as part of an International Monetary Fund program. In the market turmoil of 1997 to 1998, the Czech Republic's exchange rate target became untenable, and it adopted IT. To fit the regime to postcommunist transition, the Czechs have introduced a net inflation index, which excludes administered, or controlled, prices. Poland and Hungary followed, trying to stabilize their inflation to better integrate with Europe. South Korea began its policy under considerable duress after its financial crisis in the late 1990s, whereas Israel was trying to lock in its already completed disinflation.

Though the list of IT countries is 26-strong, the so-called G3—the U.S. Federal Reserve, the ECB, and the Bank of Japan—are notably absent. This is especially ironic given the Fed's long, (mostly) successful history of pursuing price stability and the enthusiasm for IT of its chairman, Ben Bernanke, during his distinguished stint as an academic economist. Despite the lack of a formal commitment to a clear inflation target, there has arguably been a progression toward the IT framework and greater transparency. In February 1981, the Supreme Court defended the Fed's right to secrecy, saying, "At bottom, the FOMC [has] concluded that uncertainty in the monetary markets best serves its needs" (*Merrill v. Federal Open Market Committee*). By 2004, Bernanke described an "ideal world" in which "the Federal Reserve would release to the public a complete specification of its policy rule, relating the FOMC's target for the federal funds rate to current and expected economic conditions, as well as its economic models, data, and forecasts." Presently, this ideal has not been reached.

Along with increasing its transparency, the Federal Reserve has been edging toward a de facto target range. In 2002, the term *comfort zone* appeared in a *New York Times* headline (Stevenson, 2002) and has subsequently been used by the Fed governors themselves to describe their desired inflation level. Still, they seem to lack credibility in this target, as admitted by Bernanke (2004): "Today

**Table 37.1** Inflation Targeting—Time of Adoption and Target Range

| *Country* | *Adoption* | *Target* | *Country* | *Adoption* | *Target* | *Country* | *Adoption* | *Target* |
|---|---|---|---|---|---|---|---|---|
| New Zealand | 1990 | 1%–3% | Chile | 1999 | 2%–4% | Hungary | 2001 | 3%(±1%) |
| Canada | 1991 | 1%–3% | Columbia | 1999 | 2%–4% | Peru | 2002 | 2%(±1%) |
| United Kingdom | 1992 | 2% | Switzerland | 2000 | <2% | Philippines | 2002 | 4%–5% |
| Sweden | 1993 | 2%(±1%) | South Africa | 2000 | 3%–6% | Indonesia | 2005 | 6%(±1%) |
| Australia | 1993 | 2%–3% | Thailand | 2000 | <3.5% | Romania | 2005 | 4%(±1%) |
| Israel | 1997 | 1%–3% | S. Korea | 2001 | 3%(±1%) | Slovakia | 2005 | 3%(±1%) |
| Czech Rep. | 1998 | 3%(±1%) | Mexico | 2001 | 3% | Turkey | 2006 | 4%(±2%) |
| Poland | 1998 | 2.5%(±1%) | Iceland | 2001 | 2.5%(±1.5%) | Ghana | 2007 | <10% |
| Brazil | 1999 | 4.5%(±2%) | Norway | 2001 | 2.5% | | | |

SOURCE: Roger and Stone (2005) and national bank Web sites.

long-term inflation expectations in the United States remain in the vicinity of 2-1/2 to 3 percent, above the range of inflation that many observers believe to represent the FOMC's implicit target."

Although much of the Fed's attention during the later part of 2008 was dedicated to the ongoing financial crisis, the minutes of the December 16, 2008, Federal Open Market Committee meeting reveal the participants' deliberation whether "a more explicit indication of their views on what longer-run rate of inflation would best promote their goals of maximum employment and price stability." The financial market turmoil drove the nominal interest rate toward its lower limit of zero. An explicit positive target rate of inflation may be needed to keep the nominal interest rate positive in the wake of large negative demand disturbances, such as a financial crisis. Matt Klaeffling and Victor Lopez Perez (2003), for instance, find that the presence of the zero bound implies that the optimal rate of inflation should be somewhat higher than in its absence, even if higher long-term inflation generates stronger macroeconomic volatility.

The ECB also does not consider its policy to be an explicit inflation target but has announced that an inflation rate below 2% is desirable. It promises to direct its policy to protect this bound. The legal basis of the ECB is also similar to an inflation target, as Article 105(1) of the treaty establishing the European Community mandates price stability as "the single monetary policy for which [the ECB] is responsible." But, in keeping with the flexibility under IT, its Web site notes that the ECB "typically should avoid excessive fluctuations in output and employment" (ECB, n.d.). The de facto similarity to a flexible IT regime leads some commentators to discuss the ECB as if it were one, although its own leadership rejects the label.

Arguing from the position of an ECB board member, Jürgen Stark (2007) suggested that part of the reason why the ECB does not explicitly join the ranks of inflation targeters is in order to preserve institutional continuity with its predecessors. The new authority wishes to be seen as inheriting the German Bundesbank's tenacity in controlling inflation. With its current strategy, Stark argues, the ECB has more flexibility in medium-term and long-term objectives and in communication strategy, as it can put less emphasis on forecasts. With a firm record of keeping its informal target, perhaps the ECB has already attained the credibility that an IT regime hopes to create.

Many early adopters of IT have been quite successful under the new monetary regime, but it has proven difficult to take a definitive stance on its overall effect. Table 37.2 compares the summary statistics for measures of inflation and output growth for IT and non-IT countries. In the 10 years prior to IT, New Zealand averaged 10.8% inflation with a standard deviation of greater than 5%, 1.8% output growth with a standard deviation of 1.85%. From 1990 to 2006, inflation was below 2% with a standard deviation of about 1%, and growth was higher and less volatile, about 3% with a standard deviation of 2%. Similarly, since Canada stabilized its inflation target

**Table 37.2** Real Output Growth and Inflation Volatility for Selected IT and Non-IT Countries

| | | *Pre-Inflation Target /1982–1992* | | | | *Post-Inflation Target / 1993–* | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Avg. growth* | *St. dev. growth* | *Avg. inflation* | *St. dev. inflation* | *Avg. growth* | *St. dev. growth* | *Avg. inflation* | *St. dev. inflation* |
| Inflation targeters | New Zealand | 1.89% | 1.85% | 10.08% | 5.24% | *2.99%* | *2.02%* | *1.99%* | *0.99%* |
| | Canada | 2.84% | 2.57% | 5.09% | 2.58% | *2.81%* | *1.87%* | *1.97%* | *1.15%* |
| | United Kingdom | 2.42% | 1.95% | 5.82% | 1.45% | *2.89%* | *0.66%* | *2.51%* | *0.59%* |
| | Sweden | 1.81% | 1.72% | 6.99% | 2.47% | *2.71%* | *1.80%* | *1.70%* | *0.94%* |
| | Australia | 2.76% | 2.59% | 5.99% | 2.94% | *3.65%* | *0.90%* | *2.60%* | *1.55%* |
| Non-inflation targeters | United States | 3.02% | 2.44% | 3.52% | 1.05% | *3.21%* | *1.11%* | *2.15%* | *0.58%* |
| | Japan | 3.72% | 1.73% | 2.00% | 0.87% | *1.21%* | *1.39%* | *–0.74%* | *0.79%* |
| | Denmark | 2.27% | 1.74% | 4.68% | 2.44% | *2.34%* | *1.53%* | *1.94%* | *0.69%* |
| | France | 2.32% | 1.17% | 5.10% | 3.26% | *1.97%* | *1.21%* | *1.50%* | *0.60%* |
| | High-income countries | 3.02% | 1.40% | 4.63% | 1.55% | *2.58%* | *0.83%* | *2.49%* | *0.39%* |

SOURCE: World Bank (2008).

in 1995, inflation has been in its target 1% to 3% range almost continually, and its business cycles have been relatively mild.

Even countries that have missed their inflation targets, as have many of the adopters in developing countries, have had relatively strong performances and have reformed their policy practices in the aftermath of IT adoption. The Czech Republic is a good example of this, as it severely undershot its target in 1998 and 1999. It later began targeting headline CPI, rather than a restricted core, and made the target more strict, by promising to hold it continuously rather than just at year's end. In its first decade of IT, Czech output has been stable, averaging 4.1% with a standard deviation of 1.9%, and its inflation has been low and stable, averaging 2.1% with standard deviation 1.8%. Even when countries overshot their inflation targets, as Israel in 2002, the IT policy makers have been able to regain control. Despite very large exogenous shocks, the inflation spike was brief and has been within the target since, while real output growth has averaged over 5%.

Several attempts have been made to systematically study countries before and after implementing IT and to compare them to non-IT countries. This is complicated because in the 1980s, just before most targets were adopted, inflation and output performance was generally poor and volatile, while the 1990s were relatively prosperous and stable. It is difficult to isolate the IT effect from generally favorable world economic conditions in the experimenting countries. Laurence Ball and Niamh Sheridan (2005) use a panel regression approach with 20 Organisation for Economic Co-operation and Development countries, but find that an IT regime does not significantly improve the level or variance of output growth or inflation when they control for pre-IT conditions, which were generally worse in the adopting countries. However, Marco Vega and Diego Winkelried (2005), using a similar methodology but including developing countries into the sample, show that the policy regime choice *does* have a statistically significant effect. David Johnson (2002) finds that IT significantly reduced inflation expectations as measured by the average among professional forecasters. Frederic Mishkin and Klaus Schmidt-Hebbel (2007) find that IT reduces inflation levels and volatility and lowers interest rates, and this effect is stronger in industrial countries. More conclusive evidence about IT may only come with more practical experience and further academic effort.

## Conclusion

The adoption of IT may very well be the most significant development in monetary economics over the last couple of decades. The number of central banks around the world that have officially adopted IT as the framework for their conduct of monetary policy has been steadily increasing. Although the time frame for evaluating the success of IT may be relatively short, the preliminary results that have emerged are encouraging. None of the IT adopters have abandoned this policy, and several central banks that have not formally announced IT as their policymaking framework have acted in a way that is fully consistent with it.

Although little theoretical literature on IT existed prior to New Zealand's adoption of IT in 1990, the policy has been studied extensively in academia and in the central banks' research departments. There may not be overwhelming consensus on the exact operational details of implementing IT in practice, but the general opposition to IT is rather slim. Given the international monetary trends in the past few decades, it seems reasonable to suggest that IT will continue to be the predominant paradigm for the conduct of monetary policy in the near future.

## Notes

1. In a letter to the editor, Friedman said the full quote was, "In one sense, we are all Keynesians now; in another, nobody is any longer a Keynesian."
2. Svensson (2008) suggests that the Bundesbank, although declaratively a monetary aggregate targeter, in fact followed IT through the 1980s, which may explain its success in achieving inflationary stability earlier that other central banks.
3. A pricing mechanism due to Calvo (1983) or price adjustment costs are popular theoretical devices that motivate these pricing decisions.

## References and Further Readings

Ball, L., & Sheridan, N. (2005). Does inflation targeting matter? In B. Bernanke & M. Woodford (Eds.), *The inflation-targeting debate* (NBER Studies in Business Cycles No. 32, pp. 249–276). Chicago: University of Chicago Press.

Bernanke, B. S. (2004, January 13). *Remarks by Governor Ben S. Bernanke.* Speech at the Meetings of the American Economic Association, San Diego, California. Available from http://www.federalreserve.gov/boarddocs/speeches/2004/200401032/default.htm

Bernanke, B. S., Laubach, T., Mishkin, F. S., & Posen, A. S. (1999). *Inflation targeting: Lessons from the international experience.* Princeton, NJ: Princeton University Press.

Bernanke, B. S., & Woodford, M. (Eds.). (2006). *The inflation-targeting debate* (NBER Studies in Business Cycles No. 32). Chicago: University of Chicago Press.

Board of Governors of the Federal Reserve System. (2008, December 15–16). *Minutes of Federal Open Market Committee.* Available from http://www.ircforex.com/news/726-minutes-federal-open-market-committee

Calvo, G. (1978). On the time consistency of optimal policy in a monetary economy. *Econometrica, 46*(6), 36–48.

Calvo, G. (1983). Staggered prices in a utility maximizing framework. *Journal of Monetary Economics, 12*(3), 383–398.

Clarida, R., Gali, J., & Gertler, M. (1999). The science of monetary policy: A New Keynesian perspective. *Journal of Economic Literature, 37*(4), 1661–1707.

European Central Bank. (n.d.). *Objective of monetary policy.* Available from http://www.ecb.int/mopo/intro/objective/html/index.en.html

Friedman, B. (2004). Why the Federal Reserve should not adopt inflation targeting. *International Finance, 7*(1), 129–136.

Friedman, M. (1968). The role of monetary policy. *American Economic Review, 58*(1), 1–17.

Greenspan, A. (2004). Risk and uncertainty in monetary policy. *American Economic Review Papers and Proceedings, 94*(2), 33–40.

Johnson, D. (2002). The effect of inflation targeting on the behavior of expected inflation: Evidence from an 11 country panel. *Journal of Monetary Economics, 49*(8), 1521–1538.

Klaeffling, M., & Lopez Perez, V. (2003). *Inflation targets and the liquidity trap* (European Central Bank Working Paper No. 272). Frankfurt, Germany: European Central Bank.

Kydland, F., & Prescott, E. (1977). Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy, 8*(49), 473–493.

Merrill, D. R., v. Federal Open Market Committee of the Federal Reserve System. (1981, February). *Defendant's supplemental memorandum of points and authorities* (United States District Court for the District of Columbia, Civil Action No. 75-0736).

Mishkin, F. S., & Schmidt-Hebbel, K. (2007). *Does inflation targeting make a difference* (NBER Working Paper No. 12876). Cambridge, MA: National Bureau of Economic Research.

Roger, S., & Stone, M. (2005). *On target? The international experience with achieving inflation targets* (IMF Working Paper WP/05/163). Washington, DC: International Monetary Fund.

Rogoff, K. (1985). The optimal degree of commitment to an intermediate monetary target. *Quarterly Journal of Economics, 100*(4), 1169–1189.

Stark, J. (2007, January 19). *Objectives and challenges of monetary policy: A view from the ECB*. Speech at a conference on inflation targeting, Magyar Nemzeti Bank, Budapest, Hungary.

Stevenson, R. W. (2002, August 4). The Fed's evolving comfort zone. *The New York Times,* p. B4.

Svensson, L. (1998, Winter). Monetary policy and inflation targeting. *NBER Reporter,* pp. 5–8.

Svensson, L. (2002). Inflation targeting: Should it be modeled as an instrument rule or a targeting rule? *European Economic Review, 46*(4–5), 771–780.

Svensson, L. (2003). What is wrong with Taylor rules? Using judgment in monetary policy through targeting rules. *Journal of Economic Literature, 41*(2), 426–477.

Svensson, L. (2008). *What have economists learned about monetary policy over the past 50 years?* (Sveriges Riksbank Press Release No. 39). Stockholm: Sveriges Riksbank.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy, 39*(1), 195–214.

Thiessen, G. (1999, March 11). *Then and now: The change in views on the role of monetary policy since the Porter Commission.* Speech at the C. D. Howe Institute, Toronto, Ontario, Canada.

Vega, M., & Winkelried, D. (2005). Inflation targeting and inflation behavior: A successful story? *International Journal of Central Banking, 1*(3), 153–175.

Woodford, M. (2003). *Interest and prices: Foundations of a theory of monetary policy.* Princeton, NJ: Princeton University Press.

World Bank. (2008). National accounts and prices statistics. In *World development indicators 2007*. Available from http://go.worldbank.org/3JU2HA60D0

# 38

# Debates in Macroeconomic Policy

John Vahaly
*University of Louisville*

Modern economics began with Adam Smith's *The Wealth of Nations* (1776), which established the promotion of rising living standards as the major economic goal. Smith labeled economic thought of his time as Mercantilism, which seemed to place emphasis on the acquisition of money, gold and silver primarily, as the key to prosperity.

*The Wealth of Nations* integrated eighteenth-century political and economic thinking to present a new framework for economic organization to promote economic well-being. Smith demonstrated that the then prevailing emphasis of government accumulation of specie—gold and silver—was misplaced. For Smith and the following classical school, the correct view of economic progress had nothing to do with money.

Just as in modern views of economic growth, rising living standards depended on increasing the productivity of resources, primarily labor and physical capital. Many factors contribute to rising levels of productivity with the classical tradition favoring reliance on private entrepreneurial activity with essential but limited roles for government.

Smith understood that market or capitalist economies require government oversight and productive contributions. Private markets can fail to operate in society's interests in a variety of ways, requiring government intervention to reallocate resources as with public goods and externalities and to promote competition where impediments exist. Smith also understood the income distribution or fairness issue and was an early supporter of progressive taxation. However, with respect to the business cycle with contractions and expansions of national production, Smith and the classical economists supported a laissez-faire approach. Thus, if a recession arose, excess output could always be sold through price reductions. Similarly, workers could always find work through wage declines. A general deflation would then gradually move the economy back to its full-employment resting place. Flexible wages and prices would move in the opposite direction, should an economy become overextended in an inflationary boom

This leads to the second tenet of classical macroeconomics—the relevant time frame of interest is the long run. Whatever short-term instability existed was best remedied by the economy's self-correcting mechanism. The more quickly wages and prices adjusted to economic shocks, the faster the economy would return to its full-employment level of income and output.

Finally, only the supply or production side of the economy mattered for economic growth. Say's law of markets, namely, that "supply creates its own demand," meant the demand side was not a concern. Employment of resources automatically generated incomes sufficient to purchase what was produced. Any income not spent was automatically placed back into the economy through business investment spending on the physical capital through the loanable-funds market.

The nineteenth century is also associated with establishment of a global trading system that relied on both free trade and the "classical" gold standard. The case for trade was inherent in Smith's discussion of specialization and exchange. However, the formal logic for trade was provided by David Ricardo's principle of comparative advantage. The idea that competition or trade between nations raises living standards remains a principle in modern economics.

The international gold standard served two functions. The first was to create an international payments system

to finance trade. The second was to regulate the amount of money in circulation. In a true gold standard, gold circulates both as money and as a commodity and is freely exchangeable between such uses. Because the amount of money was determined by a commodity, the price level was highly correlated with changes in the amount of gold in circulation. Prices rose and fell with new gold discoveries, bullion ship sinkings, and new technologies in gold mining. Such situations produced alternate inflations and deflations as nations allowed their domestic money supplies to be determined by the international gold standard. This system provided two major benefits: it was self-regulating and did not require direct monetary intervention, and it produced long-run price stability. The short-term record, however, was one of instability and unpredictable swings in prices.

David Hume is credited with formalizing this link between the money supply and the price level in the quantity theory of money, MV = PY. The money supply, M, multiplied by its annual turnover rate or velocity, V, equaled nominal gross domestic product (GDP), PY. Nominal GDP has two components: the level of prices, P, and real GDP, Y, calculated using prices from a base year. The quantity theory was famously used to explain how a mercantilist country that attempted to block imports, forcing trading partners to pay for exports with gold, must fail. A simple interpretation of the quantity theory shows that an increase in gold, thought to be the source of wealth, puts upward pressure on P, the domestic price level, thereby reducing exports while promoting imports. In the long run, rigging trade to acquire gold must fail. Better policies were free trade, private-market resource allocation, and laissez-faire.

During the nineteenth century, the U.S. economy operated according to classical rules. In the 1890s, the United States experienced several major recessions. Real output actually fell 3 years with the price level declining in 6 years. Real output was higher at the end of the decade by nearly half, but the short-run economic record saw recession and deflation. For macroeconomic thinking, this was how the business cycle should operate.

## The Panic of 1907 and Creation of the Federal Reserve

The classical hands-off approach also applied to commercial banking and financial markets. Twice, the United States had established a central bank prior to the twentieth century, but neither survived original chartering. In 1907, as had previously happened, a major New York bank failed. This triggered a financial panic that had the impact of producing a sharp drop in the U.S. money supply and subsequent reduction in real output. The panic would have been worse had not John Pierpont Morgan opened his personal fortune to deposit money in banks in a vote of financial support. A century later, a similar but larger crisis would confront the U.S. economy.

The lesson drawn from this episode was to make clear the need for a central source of reserves for the U.S. banking system. What emerged was the Federal Reserve Act (1913) and the Federal Reserve System that opened in 1914. The System, composed of 12 regional banks, was dominated by the New York Federal Reserve until events in the Great Depression resulted in an overhaul of the U.S. monetary structure.

## John Maynard Keynes and the Great Depression

No event in U.S. economic history has had more impact than the Great Depression. Debate continues today as to the causes and consequences of this economic disaster. The collapse that lasted from August 1929 to March 1933 saw real output fall more than 26%. Unemployment rose from 3.5% to over 25% while the price level dropped 26%. The breakdown appeared to mirror past macroeconomic declines in that falling output and rising unemployment were accompanied by deflation. However, this decline seemed to have no end. And at Cambridge University, a new approach to the business cycle was being created to help deal with a global depression.

In *The General Theory of Employment, Interest and Money* (1936), John Maynard Keynes revolutionized macroeconomics. Classical thinking was discarded as emphasis was changed from the long to the short run and from supply to demand. For Keynes, events in the 1930s required immediate policy response. Furthermore, economies had excess supplies of labor, capital, and other productive resources. Therefore, the key problem was a lack of short-term demand or spending on output. The key question became how to best stimulate total expenditure on GDP. The private sector was in disarray as spending by consumers (C) and businesses (I) fell between 1929 and 1933. Consumer expenditure dropped from $752 to $558 billion (using prices from the year 2000) with private investment spending collapsing from $91 to just $17 billion over this period.

The Keynesian alternative outlined in *The General Theory* called for the use of deficit spending as a short-term policy to restore full employment. The federal government could either directly raise spending (G) or reduce taxes (T), which through a multiplier process would ultimately raise output and income. Alternatively, the Federal Reserve could pursue expansionary policies that would help lower interest rates to promote new spending. Following publication of *The General Theory,* however, fiscal policy became the preferred stabilization policy tool. The alternative to this was to wait an indeterminate time period while the economy's self-correcting forces operated. While U.S. monetary and

fiscal authorities took action, unemployment stayed in double digits until the outbreak of World War II. The modern debate over the government's role in stabilizing the economy had started.

Keynes offered an alternative for ending the Depression. But what had caused the collapse? A recession had started in the summer of 1929, which was soon followed by the Great Crash during the fall of 1929. The negative impact of the 1930 Hawley-Smoot tariff increase contributed. Perhaps the two greatest economists of the twentieth century, John Maynard Keynes and Milton Friedman, both agreed the greatest problem for the United States was a series of policy mistakes made by the Federal Reserve. The Fed failed to contain a series of financial panics between 1930 and 1933. The financial system was so unsound that the newly elected Franklin Roosevelt had to close all banks to help restore financial soundness. This overhaul was accompanied by new laws guaranteeing depositor money (Federal Deposit Insurance Corporation) and a new structure for the Federal Reserve itself in which the Board of Governors and the Federal Open Market Committee (FOMC) were placed in charge of monetary policy and the determination of credit conditions and interest rates. These actions were taken after 1933, after both the nominal money supply and the number of U.S. banks had dropped by one third.

## The Postwar Study of Economic Growth

With the Depression behind them, macroeconomists again took up the study of economic growth. Logically, these efforts took Keynesian ideas and developed their implications for growth theory. Evsey Domar and Roy Harrod independently created similar models of economic growth that were familiar to Keynesian economists. The Harrod-Domar growth model emphasized the importance of saving, a major problem for low-income countries, as well as the productivity of the capital stock.

Robert Solow added a fundamental contribution to growth theory that remains the starting point for such analysis. Solow's model, referred to as the *neoclassical growth model,* was a supply-side approach to thinking about raising living standards. For the next decade, economic growth theory built on Solow's model, which gave new life to the classical interest in long-run growth.

Solow's theory contained two insights. Starting with the simplest model of supply, Solow showed that *rising* living standards, the goal of economic growth, do not arise from having a relatively high *level* of national saving. Nations can divert current income to saving and then to investment or real capital formation. This will raise living standards, but only once. Additional saving that can provide workers more capital (machines) with which to work can raise productivity levels. However, additional saving must occur if workers are to become even more productive. In Solow's model, technological change is the reason why nations can experience continuous improvements in real living standards. When technological advance occurs, workers in affected sectors can become more productive. Modern economic growth is thus a record of technological change.

Solow's model makes no attempt to model or to explain a theory of technological change. The neoclassical model assumes that technology changes for reasons that cannot be predicted, or that technological change is exogenous and independent of the model. Somewhat surprisingly, the key to growth is not explained. The Solow model nonetheless remains the starting point for the economic growth studies.

## The Arrival of Keynesian Macropolicy in the United States

Keynesian macroeconomics came to the United States in the 1960 presidential election. A recession was under way in 1960, and John F. Kennedy proposed a deficit-financed tax cut to stimulate the economy and end the recession. The Kennedy tax plan was not enacted until 1964 and was accompanied by two contemporaneous fiscal shocks. The first came from the new President Lyndon Johnson's Great Society program and its War on Poverty. The second was the Vietnam War. These three demand shocks produced an overheated economy that by the end of the decade had produced a very low unemployment rate of 3.5% but an accelerating inflation rate that topped 5%. Two points stood out to economists in 1969. First, there had been no recessions during the 1960s, and Keynesian-style demand management was given credit for the longest period of economic expansion in U.S. history up to that time. Second, the economy was behaving precisely as had been expected. Excess demand had resulted in higher prices, or the increase in inflation, while driving down unemployment. In macroeconomic books, the Phillips curve inflation–unemployment trade-off emerged as the tool economists used to explain how the U.S. economy operated.

The ascendancy of the original Keynesian model was challenged by macroeconomic events of the 1970s. The decade began with a recession that produced higher unemployment but without any decline in inflation. This surprised Keynesian economists, who had come to rely on the Phillips curve to portray an inverse relationship between inflation and unemployment. In the early 1970s, however, the trade-off seemed to be weakening. Then President Nixon resorted to wage and price freezes and controls in 1971 to slow the 5% inflation, then viewed as a major macroeconomic problem. President Nixon, remembering his loss to Kennedy in 1960, did not want to risk another recession in 1972, a presidential election year. While the U.S. experiment with wage and price controls seemed to initially produce positive results with

inflation starting to fall, international events would overtake the attempt to slow inflation.

## Stagflation and the Reemergence of Conservative Macroeconomics

War in the Middle East in 1973 lead the Organization of Petroleum Exporting Countries (OPEC) oil cartel to quadruple crude oil prices that year. As this energy price shock spread through the global economy in 1974, it produced a phenomenon that required a new word—stagflation—to describe the simultaneous occurrence of both higher inflation and unemployment. The Phillips curve seemingly could not account for this event. For U.S. policy makers attempting to deal with what was a new macroeconomic crisis, the choices were few. Fiscal policy was effectively paralyzed by the Watergate investigation, and Fed chairman Arthur Burns was asked repeatedly by Congress and a growing number of U.S. citizens to do something about double-digit inflation and rising unemployment.

The Fed's dilemma, which also directly applies to fiscal policy action, had to do with the basics of the Keynesian model. Developed for a country suffering from very high unemployment, the Keynesian model was concerned only with short-run demand or spending issues and what government might do to make up a spending gap should private (C + I) spending be too low to produce full employment. Supply was not part of this framework. This omission was correct for the 1930s, as there were abundant quantities of labor, capital, and other scarce resources. But this omission was to undermine the Keynesian consensus in the early 1970s.

Edmund Phelps and Milton Friedman explained why the original Phillips curve—spending approach to understanding countercyclical stabilization policy—was wrong. The fatal omission was on the supply side. How would the economy respond to changes in the unemployment (output)–inflation (price level) mix as these two variables changed? The Friedman-Phelps idea was that the rising demand-side inflation in the late 1960s caused a supply-side reaction. U.S. workers saw prices rising for commodities they purchased, but without an increase in nominal wages, their real purchasing power or living standards fell. To prevent this, workers would demand a nominal wage increase at least equal to what they *expected* inflation to be. The increase in wages would prompt another round of price increases. Once set into motion, this wage–price spiral seemingly took on a life of its own. Wage increases boosted the cost of doing business, as workers received higher wages to offset inflation, thereby helping to perpetuate the upward rise in prices. All this was missing from the original interpretation of the Phillips relation.

The new Friedman-Phelps interpretation, termed the *expectations-augmented* Phillips curve, was revolutionary. It correctly predicted that the short-run trade-off between inflation and unemployment lasted only as long as the actual and expected or perceived inflation rates differed. In 1969, actual inflation was above 5%, but expected inflation, reflecting inflationary expectations, lagged. Once inflationary expectations rose to match reality, the economy had restored balance to price, or inflation, and wage growth. This is the logic that underlies the modern natural rate hypothesis that identifies *full-employment real GDP* as that rate of output where inflation tends to remain constant. Inflation tends to remain fixed at the natural rate because actual and expected inflation are the same. Workers are not expecting prices to grow faster than their wages, and managers are not expecting labor costs to grow faster than prices. Corresponding to the natural rate of real GDP is the natural rate of unemploy ment at which the economy has reached this inflation and output balance point. Milton Friedman famously said that there is no way to quantify this natural rate of unemployment, but other economists have tried to do this for the United States and other economies, and current views allow for the natural rate of unemployment to vary over time.

If the Phillips curve trade-off lasted only as long as there was a difference between actual and expected inflation, it appeared to make little sense to attempt stabilization policy that required the presence of "fooling," using Friedman's term, on the supply side. For Friedman, the short-run Phillips trade-off required workers to not correctly understand what was happening to prices. In other words, there had to be unanticipated inflation for policy makers to exploit the trade-off. In rebuttal, Keynesians identified a wide variety of both nominal and real rigidities in the economy that prevent relatively quick matching of actual and expected inflation. This debate is an ongoing area of interest and research.

Also during the early 1970s, Robert Lucas and Thomas Sergeant extended the conservative rebirth by broadening the role of "rational expectations" and going well beyond the Friedman-Phelps contribution. Friedman and Lucas provided the cornerstone for a macroeconomic counterrevolution. For both, the task of stabilizing the U.S. economy was a goal not worth pursuing.

Friedman explained that macroeconomic stabilization policy was doomed to fail because it was too difficult to achieve. Four beliefs were offered as proof. First is the challenge posed by time lags with obtaining economic information, forming policy, policy implementation, and policy effects, all of which are "long and variable." For example, the 1964 Kennedy-Johnson tax cut was enacted to help end the 1960 recession. Next is the frank admission that we do not understand the U.S. economy well enough to effectively stabilize it. Contemporary difficulties with stagflation seemed to clearly make this point. The rational expectations revolution seemed to emphatically make this point. In the Lucas critique, Lucas also pointed to the naïve

approach to understanding and modeling expectations in the original Keynesian model.

Third is the miserable past performance of stabilization efforts, particularly with monetary policy. The three major economic contractions of the twentieth century were caused or made far worse by restrictive monetary policies. Finally, there is the issue that leads to Friedman's policy recommendation—policy rules are inherently superior to discretionary policy. While the Congress and the president are elected and answer to voters, the Fed and monetary policy are effectively shielded from the ballot box.

Friedman's macroeconomic school was monetarism. To deal with the rules-discretion problem, Friedman proposed a simple monetary rule of matching monetary growth to real GDP growth. Thus, as the economy grew, the Fed would simply add enough money to the economy to finance the additional transactions. During the 1970s, faced with rising inflation, this type of approach made sense to a growing number of people. Friedman (1992) famously stated that "inflation is everywhere and always a monetary phenomenon" (p. 193), and his monetary rule was based on this view. In essence, Friedman proposed lowering inflation, or the growth of prices, by matching money growth to no more than an expanding economy needed to finance additional transactions. Inflation arose when "too many dollars, or too much spending, chased too few goods." Monetarism dealt directly with the too many dollars issue, leaving the supply-side, or "too few goods," to the private economy where it belonged. For Friedman, all the Fed had to do was add sufficient reserves to the U.S. banking system to allow the money supply to expand as fast as real output. People had known for centuries that excess creation of money by resorting to the printing press could produce hyperinflation, and modern examples of this can still be found. On the other hand, the Depression clearly demonstrated the consequences of large monetary reductions.

Friedman's constant growth rate rules (CGRR) offered a secure return to the laissez-faire macroeconomics of the past. Its popularity was ended, however, in the early 1980s with the deregulation of U.S. commercial banking and financial markets. The high inflation of the 1970s and the new opportunities for managing cash holdings created in the early 1980s had a major impact on the demand for money.

In the GDP version of the quantity theory, MV = PY. Expressed as rates of growth, the quantity theory is written as follows: money growth + velocity growth = inflation + real GDP growth. Friedman had linked money growth to real GDP growth, assuming the growth of both V and P would be modest at best. If V is interpreted as the amount of money people choose to hold as a fraction of nominal income, PY, then instability in the growth of velocity ruins the CGRR. If velocity grows or declines in an unpredictable fashion, a fixed monetary growth policy would be destabilizing.

Robert Lucas and Thomas Sergeant had prepared the way for another attempt to restore laissez-faire macroeconomics while attempting to increase understanding of the U.S. economy. It was first common to refer to their approach as the *rational expectations school.* The label soon changed to the *new classical school,* both to reflect the ties back to pre-Keynesian economics that placed emphasis on long-run supply issues and to indicate that there was something new being studied.

The Lucas approach was pioneering in that it appeared to restore the primacy of microeconomics in the effort to comprehend macroeconomic aggregates. In microeconomics, economic agents—consumers and producers—are modeled as being rational decision makers. Consumers maximize utility; producers maximize economic profit. Macroeconomics should reflect this microfoundation. The theory of the business cycle that first emerged from this approach was initially quite similar to other macroeconomic schools. In Friedman's natural rate hypothesis, demand disturbances had caused workers to be "fooled" when demand-side inflation pushed prices above wages. In the new classical school, the presumption was the divergence that occurred between what economic agents expected prices, or inflation, and actual values arose because of "expectational errors." More formally, explanations of the business cycle presumed that there would be "price surprises," as happened in the late 1960s, for the economy to move away from its natural rate of output.

The logical conclusion to draw from this approach was that attempts to stabilize the economy could not be productive. Successful stabilization policy would have to generate a price surprise for it to matter for the real economy, but rational economic agents will always be able to divine macropolicy direction. The national economy would operate more effectively if stabilization policies were eliminated. Monetary policy would also be directed to operate more or less along Friedman's lines so that there would be no surprises that would contribute to macroeconomic instability.

The new classical school added microeconomic rigor to the debate about macroeconomic stabilization policy. As with monetarism, it ran into difficulties. Few economists had difficulty accepting the idea that expectations about macroeconomic variables, especially inflation and nominal interest rates, were formed rationally using available information. However, the claim that monetary and fiscal policy could not affect output and employment in the short run seemed incorrect. As mentioned earlier, Keynesians pointed to a variety of potential impediments to price and wage adjustments that delay movement from one equilibrium to the next. There may indeed be no surprise in the Fed's actions with the money supply, but those actions will have real impacts in the macroeconomy. In macroeconomic jargon, money mattered and was not "neutral" in the short run. In the long run, the quantity of money does not matter for living standards, excluding small gains from

avoiding barter. The quantity of money does affect nominal magnitudes, prices, nominal wages, nominal interest rates, and nominal GDP. It does not affect real macroeconomic values. The new classical school seemed to be saying that it was also neutral in the short run unless some type of surprise occurred.

The idea that there must be uncertainty about future actions had unexpectedly widespread impacts in macroeconomics. This introduced the idea that game theory could be applied to the decision made concerning monetary policy. One of the difficulties of the 1960s and 1970s was that there was little coherence in actions taken by the Federal Reserve when economic conditions deteriorated. Improved policy response required actions to be credible and consistent with no attempts to pursue conflicting goals. This was clearly a step beyond the original promises of fine-tuning made by economists in the 1960s. This issue became especially important in the debate over the U.S. macroeconomic policy in response to the collapse of the U.S. housing and financial markets in 2009.

## The Great Inflation and the Volcker Recession: Disinflation in the 1980s

In 1979, OPEC pushed crude oil prices over $30 per barrel, producing the second energy price shock of the decade. The economy slowed as prices rose even faster with a minor recession in 1980. That was also a presidential election year, and voters wanted to know what was to be done to fix the economy. President Carter had just appointed a new Fed Chair, Paul Volcker, in 1979, with a clear order to bring double-digit inflation under control. In 1980, some measures of inflation were approaching 15% with interest rates moving over 20%. Ronald Reagan's approach to the inflation problem presented a new approach to the stagflation problem. If supply problems had created the difficulties, supply-side policies had to be the remedy.

The supply-side school that emerged in the early 1980s shared the laissez-faire philosophy of the more academic monetarists and new classical schools. This was interpreted to lend support to federal income tax rate reductions as a means to lower inflation. The logic was that lower income tax rates would increase work incentives and therefore supply. Because part of the inflation problem is "too few goods," the tax cut was an anti-inflationary program. The Reagan platform also included expansion of government spending as part of a defense buildup. Added with the income tax reductions, this led many economists to conclude that fiscal expansion was being adopted that would only add to current inflationary pressures.

President Reagan followed through on his electoral program: Defense spending rose while federal taxes were lowered by the Kemp-Roth tax cut. Paul Volcker and others at the Fed tried to make clear that fiscal policy was too expansionary. In 1981, the Fed acted to slow total spending in the economy, thereby creating the worst recession since the Depression. Unemployment was pushed over 10% in some months in 1982, but the Fed's commitment to lower inflation seemed to take hold. By 1983, the inflation rate fell from 12% to 4%, adding the term *disinflation* to the U.S. economic vocabulary. Faced with no alternative, the Fed had to slow the economy sufficiently to reduce both actual and expected inflation. In the spirit of the rational expectations movement, this meant that the Fed had to make clear its commitment to restrictive policies until its inflation target was achieved.

## Real Business Cycle Models

The new classical school had based its price surprise model on concepts imported from microeconomics, including near-perfect flexibility of wages and prices. While this is a standard assumption in microeconomics, it appeared to many economists that it is demonstrably not the case in macroeconomics. The new classical assumption of wage and price flexibility was necessary for its business cycle model to operate as its supporters explained it. When this assumption was challenged, along with its prediction of short-run monetary neutrality, the new classical school changed.

Edward Prescott moved past the new classical school in the early 1980s when he combined the Lucas rational expectations approach with evidence from the economy's growth record. In Solow's neoclassical model, higher real GDP per capita arose from capital accumulation and technological change. Any growth in living standards that could not be accounted for by increases in the capital and labor must be due to exogenous or autonomous change in technology. This approach of measuring technological change as a residual from a basic growth equation produced a surprising result. Solow's technology residual seemed to be highly correlated with detrended changes in real GDP. Could technological change, a supply-side phenomenon, be the source of the business cycle? If this were true, and if strict microfoundations applied, then all government stabilization effort should be focused on promoting creation and application of technological change, understood to be a relatively broad concept.

This new real business cycle (RBC) model completely changed the interpretation of the business cycle. If short-run microconsumer and producer markets are continuously in equilibrium, their aggregates should also display this result. The macroeconomy is subject to external shocks, but growth evidence pointed to technology or supply-side factors as the cause. The business cycle arose from real shocks to the growth process, and the resulting cyclical disturbances must be optimal responses to these changes. Because micro–decision making is optimal, in the absence of market failure, macrodisturbances could be interpreted as movements from one equilibrium state to another.

The case for laissez-faire had seemingly been reestablished. Macroeconomics returned to its starting point. The task of raising living standards was confirmed as being a supply-side, long run (or short run, as they could be interpreted as being equivalent), with little discretionary role for government.

While the RBC model was being developed, new Keynesian macroeconomists also built on rational expectations and microfoundations, but with the goal of making stabilization policy's foundations even stronger. Distinctions are still made between the long and short run (and a medium term) in which stabilization policy efforts can be useful. For a variety of market imperfections and rigidities, both monetary and fiscal actions have short-term real impacts.

## The Savings and Loan Crisis

Macroeconomic policy debates in the 1980s covered a variety of issues, from the growth of the "twin deficits" to interest-rate targeting by the Federal Reserve. Major macroeconomic events of the 1980s included the 1981 to 1982 recession, the stock market crash on October 19, 1987, and the collapse of Savings and Loan (S&Ls) banks. S&Ls originally provided fixed, 30-year mortgage financing for the U.S. housing market. This traditional role was discarded during the economic turmoil of the late 1970s and 1980s. With the hope of fostering more competition, regulations were relaxed as S&Ls moved financial activities into new higher return but much riskier markets. The S&Ls, in a scenario to be repeated 15 years later, ignored risk in the pursuit of short-term profit opportunities. Speculative investments in commercial and residential real estate, lax and ineffective regulation, and outright fraud finally bankrupted the industry. The federal government took over the failed S&Ls in 1989 at significant taxpayer cost. A regulatory commission was established to reform and restructure both S&Ls and their regulation. S&L assets were sold off over time, but a precedent had been set. Overextension in the banking sector appeared to have a new lender of last resort: the U.S. taxpayer. In addition, federal policies now gave additional responsibility to private corporations, or government-sponsored enterprises (GESs), to support the goal of promoting homeownership. These agencies, in particular Fannie Mae and Freddie Mac, would play a major part of the next major wave of banking failures.

## The Rebirth of the Economic Growth Theory

As the liberal–conservative debate over the usefulness or even wisdom of stabilization policy was under way, macroeconomics returned to the study of economic growth and Solow's neoclassical growth model. While the new classical model's explanation of the business cycle had run into difficulties, interest in the creation of macromodels derived directly from the rational world of microeconomics expanded into the study of economic growth.

The new growth theory sought to explain changes in Solow's residual or exogenous technological change. Using the microfoundations approach, technological change was analyzed as the result of profit-maximizing business activity. Firms engage in developing new technologies, which can include development of labor skills or human capital, when economically profitable. In private markets, patent protection and government subsidies were ways in which individual firms could gain from developing new products and markets. The importance of competition—even if imperfect—trade, property rights, and other necessary factors for growth were emphasized as key underpinnings for promoting change and higher living standards. However, the goal of modern growth theory to help narrow the gap between average living standards between nations is still being pursued.

## Activist Macroeconomic Policy: The Great Depression and the Great Recession of 2007

The 1990s saw two macroeconomic milestones. The first was the "Long Boom," the longest cyclical expansion in U.S. history. This period lasted 120 months from March 1991 (trough) to March 2001 (peak). Following the 1990 to 1991 recession, the Fed followed countercyclical expansionary monetary policy, pushing the fed funds rate from a prerecession high of 8% to a 1992 low of 3%. Fed policy was even more expansionary during the 2001 recession. The fed funds rate was 6% at the beginning of 2001, but by late 2003, it was 1%. By 2009, the Fed moved even farther by lowering the target fed rate to essentially 0% following the start of the 2007 recession. In an old macroeconomic phrase, the Fed was actively "leaning against the wind."

The second was the emergence of another boom in the U.S. housing market. This boom and rising home prices encouraged financial markets to create bundles of real estate assets or mortgage-backed securities (MBSs) that allowed buyers to participate in the ongoing housing boom. Supporters of these new financial instruments thought that they diversified risks associated with investing in real estate markets. However, the opposite happened when the bubble began to collapse in 2006. Home prices began falling and home foreclosures rose markedly. MBSs and other widely marketed financial instruments became unmarketable, crippling U.S. credit markets. The economy began slowing with a recession starting in 2007 that also triggered a major retreat in stock prices. Many U.S. households saw significant reductions in personal wealth as the

value of both real estate and equity holdings were significantly lowered.

The 2007 recession posed a new set of challenges for the conduct of monetary and fiscal policies. Both the U.S. Treasury and the Fed were called on to deal with a banking crisis that was much larger than the S&L crisis but not as severe as during the Depression.

In the 1930s, fiscal policy was characterized by deficit spending arising from both the weak economy and New Deal spending programs. The federal deficit in 1934 was 5.9% of GDP, but it fell to just 0.1% in 1938. In 2009, projected deficits were on the order of 10% of GDP.

Monetary policy differences between the 1930s and 2009 were even more dramatic. Between 1929 and 1933, the nominal money supply fell by one third. This drop arose from Fed inaction, as successive waves of financial panics crippled the banking system. In 2009, the Fed was actively pursuing the opposite course through conventional monetary expansion supplemented with an array of new lending and asset-purchasing programs unprecedented in Fed history. Interest rates have been pushed to historic lows, and Fed Chair Ben Bernanke has made it clear that expansionary policy will continue until financial market health is restored.

## The Current Debate: Activism Versus Laissez-Faire

As this central debate about stabilization policy continues, the Great Depression of the 1930s continues to affect this macroeconomic debate. In the Keynesian view, problems in private spending coupled with the near collapse of the banking and financial sector produced the worst economic crisis in U.S. history. Through Keynes, a new approach was developed to mitigate the worst consequences of cyclical instability. Ironically, it had been the failure of the Federal Reserve to help the banking sector that contributed to the depth of the Depression. The possibility of such failures remains.

In the twenty-first century, these events are being studied again as economies struggle with financial, banking, and credit crises. The new consensus requires that action be taken to reduce adverse impacts arising from either demand or supply shocks. There will always be debate about how much and how far these actions should be taken. Macroevents and theoretical developments have clearly shown, however, that neither simple laissez-faire nor fine-tuning are sensible approaches to follow.

## References and Further Readings


Auerbach, A. (2003). Fiscal policy: Past and present. *Brookings Papers on Economic Activity, 2003*(1), 75–122.

Ball, L., Mankiw, N. G., & Romer, D. (1988). The new Keynesian macroeconomics and the output-inflation tradeoff. *Brookings Papers on Economic Activity, 1988*(1), 1–65.

Barro, R. J. (1991). Economic growth in a cross section of countries. *Quarterly Journal of Economics, 106*(2), 407–433.

Bernanke, B., & Mishkin, F. (1997). Inflation targeting: A new framework for monetary policy? *Journal of Economic Perspectives, 11*(2), 97–116.

Blanchard, O. (2009). *Macroeconomics* (5th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Collard, F., & Dellas, H. (2007). The great inflation of the 1970s. *Journal of Money, Credit and Banking, 39*(2–3), 713–731.

Dornbusch, R., Sturzenegger, F., & Wolf, H. (1990). Extreme inflation: Dynamics and stabilization. *Brookings Papers on Economic Activity, 1990*(2), 1–84.

Friedman, M. (1992). *Money mischief: Episodes in monetary history.* New York: Harcourt Brace Jovanovich.

Friedman, M., & Schwartz, A. J. (1963). *A monetary history of the United States, 1867–1960.* Princeton, NJ: Princeton University Press.

Keynes, J. M. (1936). *The general theory of employment, interest and money.* London: Macmillan.

Kydland, F. E, & Prescott, E. C. (1982). Time to build and aggregate fluctuations. *Econometrica, 50,* 1345–1370.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference on Public Policy, 1*(1), 19–46.

Lucas, R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics, 22,* 3–42.

Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A contribution to the empirics of economic growth. *Quarterly Journal of Economics, 107*(2), 407–437.

Melloan, G. (2009, January 13). We're all Keynesian again. *Wall Street Journal,* p. A17.

Plosser, C. I. (1989). Understanding real business cycles. *Journal of Economic Perspectives, 3*(3), 51–77.

Poterba, J. (1997). Do budget rules work? In A. Auerbach (Ed.), *Fiscal policy: Lessons from economic research* (pp. 53–86). Cambridge: MIT Press.

Sato, R. (1964). The Harrod-Domar model vs the neo-classical growth model. *Economic Journal, 74*(294), 380–387.

Smith, M. H., & Smith, G. (2006). Bubble, bubble, where's the housing bubble? *Brookings Papers on Economic Activity, 2006*(1), 1–50.

Snowdon, B., & Vane, H. R. (2005). *Modern macroeconomics.* Cheltenham, UK: Edward Elgar.

Solon, G., Barsky, R., & Parker, J. A. (1994). Measuring the cyclicality of real wages: How important is composition bias? *Quarterly Journal of Economics, 109*(1), 1–25.

Solow, R. M. (1957, August). Technical change and the aggregate production function. *Review of Economics and Statistics, 39*(3), 312–320.

Stock, J. H., & Watson, M. W. (2002). Has the business cycle changed and why? *NBER Macroeconomic Annual, 2002,* 159–230.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy, 30*(1), 195–214.

Temple, J. (1999). The new growth evidence. *Journal of Economic Literature, 37*(1), 112–156.

Tobin, J. (1964). Economic growth as an objective of government policy. *American Economic Review, 54*(3), 1–20.

# 39

# New Classical Economics

Brian Snowdon
*Durham University*

During the past 70 years, macroeconomic analysis has made enormous progress via the sustained accumulation of knowledge and is now firmly established as an essential component of effective public policymaking in every modern economy. In retrospect, it is uncontroversial that the "Keynesian revolution" transformed the approach of governments to economic management. Although the Keynesian consensus prevailed during the 1950s and 1960s, and was labeled the *neoclassical synthesis* by Paul Samuelson, the intellectual journey from Keynes's *General Theory of Employment, Interest and Money* (1936) to the contemporary consensus (the *new neoclassical synthesis*) has been marked by periods of considerable controversy (Blanchard, 2000; Woodford, 1999). In particular, during the early 1970s, there emerged a powerful intellectual counterrevolution directed against the central tenets of mainstream Keynesian macroeconomics. The ideas of this emerging school of thought, initially referred to as *rational expectations macroeconomics,* were soon labeled by Thomas Sargent the *new classical macroeconomics* (see Snowdon & Vane, 2005).

In many intellectual revolutions, there is often one person who stands out in terms of leadership. This is certainly the case in economics, where John Maynard Keynes and Milton Friedman are clear examples with respect to Keynesianism and monetarism. In the development of new classical macroeconomics, the economist who undoubtedly provided the intellectual stimulus to the new wave of ideas in macroeconomics was Robert Lucas, Jr., who in 1995 was awarded the Nobel Memorial Prize in Economics "for having developed and applied the hypothesis of rational expectations, and thereby having transformed macroeconomic analysis and deepened our understanding of economic policy" (Fischer, 1996, p. 11). In a series of highly technical papers, published during the period 1972 to 1976, Lucas established the analytical base of the rational expectations equilibrium approach to macroeconomic theorizing (Lucas, 1981). These seminal contributions, combined with the work of other prominent new classicists, such as Robert Barro, Finn Kydland, Edward Prescott, Thomas Sargent, and Neil Wallace, transformed the whole course of macroeconomic debate and led to a new classical revolution in macroeconomic analysis (Hoover, 1988).

In terms of theory, it is appropriate to divide the new classical contributions into two distinct time periods. New classical macroeconomics mark I (NCMI) evolved during the 1970s and was dominated by the development of the monetary equilibrium approach to business cycle theory (MEBCT). During the 1980s, the monetary equilibrium approach was replaced by new classical macroeconomics mark II (NCMII), more commonly known as *real equilibrium business cycle theory* (REBCT).

## Theoretical Foundations of NCMI: From Friedman to Lucas

In the late 1960s, at the height of the Keynesian consensus, Milton Friedman (1968) and Edmund Phelps (1968) provided a devastating critique of one of the most famous relationships in macroeconomics, namely, the hypothesis that there is a stable, long-run trade-off between inflation (a nominal variable) and unemployment (a real variable). The Friedman-Phelps expectations-augmented Phillips curve analysis is rightly celebrated as one of, if not *the,*

most important theoretical contributions to macroeconomic analysis in the postwar era, not least because of its influence on Robert Lucas, who extended and refined Friedman's monetarist analysis within a Walrasian general equilibrium framework (Hoover, 1988). Lucas became convinced that any modern theory of the business cycle must be consistent with the Friedman-Phelps natural rate hypothesis.

To both Friedman and Lucas the original Phillips curve is inconsistent with the so-called classical dichotomy—that is, the principles of rationality and optimizing behavior dictate that nominal variables cannot influence real variables because supply decisions should be based on relative prices. By adopting a neoclassical approach to the labor market, Friedman demonstrated that the microfoundations of the Phillips curve need to be framed in terms of real rather than nominal wages. Because the supply of labor depends on the anticipated real wage, the expected rate of inflation ($P^{e*}$) must play a key role in the determination of money wage inflation. However, once expectations are allowed to play a role alongside excess demand (proxied by unemployment as an empirical counterpart) as a determinant of money wage (or price) inflation, the idea of a unique and stable long-run trade-off is no longer theoretically plausible. Instead, there is a whole series of short-run Phillips curves, each associated with a specific expected rate of inflation, but no long-run trade-off between inflation and unemployment ($U$). In long-run equilibrium, at a natural (equilibrium) rate of unemployment—that is, a level of unemployment consistent with zero excess demand—the actual rate of inflation ($P^*$) equals the expected rate of inflation. It should be noted that the natural rate of unemployment is determined by supply-side factors and will therefore vary as supply-side conditions (such as labor market incentives) change in the economy (Friedman, 1968).

While there are subtle differences between the analyses of Friedman and Phelps, their basic idea, the expectations augmented Phillips curve, can be represented by Equation 1:

$$P^* = f(U) + \lambda P^{e*}. \quad (1)$$

Because inflation results when there is excess demand (aggregate demand exceeds aggregate supply), Equation 1 states that actual inflation depends on the level of unemployment, which captures the influence of the degree of excess demand in the economy, but is also influenced by agents' expectations of inflation. In this model, the actual rate of unemployment can be only temporarily reduced below the natural rate of unemployment by expansionary demand management policies that create *unexpected* inflation. This raises a crucial question: How are inflation expectations formed by economic agents?

Prior to the 1960s, the treatment of expectations in mainstream macroeconomics had been neglected. This changed dramatically following the incorporation, by both Friedman and Phelps, of the error-learning, or adaptive expectations, hypothesis into their analysis. According to this hypothesis, agents gradually adjust their current expectations of inflation ($P_t^{e*}$) on the basis of past rates of expected ($P_{t-1}^{e*}$) and actual inflation ($P_{t-1}^{*}$), as shown in Equation 2:

$$P_t^{e*} = P_{t-1}^{e*} + \psi (P_{t-1}^{*} - P_{t-1}^{e*}). \quad (2)$$

The acceleration of inflation during the 1970s, even though unemployment was not falling, appeared to verify the Friedman-Phelps analysis. However, to neoclassical purists, the Friedman-Phelps analysis was less than satisfactory, particularly in terms of the use of a backward-looking expectations hypothesis. Enter Robert Lucas.

## Foundations of NCMI

The new classical approach to macroeconomics, as it evolved in the early 1970s, exhibited several important features: (a) a strong emphasis on underpinning macroeconomic theorizing with neoclassical choice-theoretic microfoundations (methodological individualism); (b) a preference for embedding macroeconomic models within a Walrasian intertemporal general equilibrium framework; (c) the adoption of the key neoclassical assumption that all economic agents are *rational,* that is, continuous optimizers subject to the constraints that they face; (d) agents may have incomplete information but never suffer from money illusion; (e) only *real* magnitudes (relative prices) matter for optimizing decisions; and (f) complete and continuous wage and price flexibility ensure that markets continuously clear as agents exhaust all mutually beneficial gains from trade, leaving no unexploited profitable opportunities. Given this overall framework, the main building blocks of new classical mark I analysis comprise three key ideas, namely, (1) the rational expectations hypothesis, (2) continuous market clearing, and (3) the Lucas "surprise" aggregate supply hypothesis.

### The Rational Expectations Hypothesis

During the 1970s, the rational expectations hypothesis (REH) gradually replaced the adaptive expectations hypothesis as the dominant way of modeling endogenous expectations. A major weakness of the adaptive expectations hypotheses is the implication that economic agents make *systematic errors,* an outcome inconsistent with optimizing behavior.

In the macroeconomics literature, the REH came in two main forms. A *weak* version of the hypothesis states that, in forming expectations about the future value of a variable, rational economic agents will make the *best* (most efficient) use of *all* publicly available information about the factors that they believe determine that variable, subject to the constraint of the costs of collecting that

information. A *stronger,* and much more controversial, form of the REH is the Muthian hypothesis "that expectations since they are informed predictions of future events are essentially the same as the predictions of the relevant economic theory" (Muth, 1961, p. 316). This strong version was adopted by leading exponents of the new classical school and incorporated into their macroeconomic models. In the Muthian REH, economic agents' *subjective* expectations of economic variables will coincide with the true or objective mathematical conditional expectations of those variables. With respect to the formation of expectations of inflation ($P_t^{e*}$), the REH can be expressed as follows in Equation 3:

$$P_t^{e*} = E(P_t^* \mid \Omega_{t-1}), \tag{3}$$

where $P_t^*$ is the actual rate of inflation; $E(P_t^* / \Omega_{t-1})$ is the rational expectation of the rate of inflation subject to the information available up to the previous period ($\Omega_{t-1}$). It is important to emphasize that agents with rational expectations do not have perfect foresight. To form a rational expectation of inflation, agents will need to take into account what they believe to be the "correct" macroeconomic model of the economy. Agents will also make errors in their forecasts due to incomplete information. However, with rational expectations, agents' expectations of economic variables on average will be correct. Furthermore, unlike the adaptive expectations hypothesis, where agents make repeated errors, rational agents will not form expectations that are systematically wrong (biased) over time. If expectations turn out to be systematically wrong, agents will learn from their mistakes and change the way they form expectations, thereby eliminating systematic errors. More formally, the strong version of the REH implies that the forecasting errors from rationally formed expectations will be random with a mean of zero, will be unrelated to those made in previous periods (serially uncorrelated over time), and will have the lowest variance compared to any other forecasting method. Notwithstanding the many criticisms launched against the REH, during the 1970s, there was undoubtedly a "rational expectations revolution" in macroeconomics to such an extent that by the late 1970s, the REH was being incorporated into "new" Keynesian models with sticky wages and prices (Gordon, 1990).

## Continuous Market Clearing

A second key feature of new classical models is the adoption of the assumption that all markets in the economy continuously clear so that the economy is viewed as being in a permanent state of (short- and long-run) equilibrium. In new classical theories, equilibrium is interpreted to mean that all economic agents within a market economy have made choices that optimize their objectives subject to the constraints that they face, *including imperfections of information.* Because even rationally formed expectations can turn out to be incorrect due to incomplete information, this means that, at least until agents acquire more accurate information, a currently observed market clearing equilibrium may differ from a full information equilibrium.

The assumption of continuous market clearing is probably the most critical and controversial assumption underlying new classical analysis. The assumption of perfectly flexible prices and wages is in stark contrast to Keynesian economics, where it is assumed that, in the short run, markets fail to clear because of the slow adjustment of prices or wages, or both (Gordon, 1990; Keynes, 1936; Tobin, 1980).

## The Lucas Aggregate Supply Hypothesis

The new classical approach to aggregate supply derives mainly from Lucas (1972, 1973). In these two papers, Lucas assumes that economic agents know the current price of their own good or service, but acquire information about prices in all other markets only after a time lag. Therefore, all agents are faced with a *signal extraction* problem, in that they have to distinguish between variations in relative and absolute prices. Rational agents, seeking to maximize their utility or profits, should supply more goods, services, or labor only in response to a real (relative) rise in their supply price. A purely nominal price change—that is, an increase in price that is common in all markets—does not require any positive supply response. According to Lucas, the greater the variability of the general price level, the more difficult it will be for any economic agent to extract a correct signal from a change in prices, and hence the smaller the supply response is likely to be to any given change in prices (Lucas, 1977).

The analysis of the behavior of individual agents in terms of the supply of both labor and goods has led to what is referred to as the *Lucas surprise* aggregate supply function, in the following form:

$$Y_t = Y_{Nt} + \alpha\,[P_t - E(P_t / \Omega_{t-1})]. \tag{4}$$

Equation 4 states that output ($Y_t$) deviates from its natural level ($Y_{Nt}$) only in response to deviations of the actual price level ($P_t$) from its (rational) expected value [$E(P_t / \Omega_{t-1})$]—that is, in response to an unexpected (surprise) increase in the price level. For example, when the actual price level turns out to be greater (less) than expected, individual agents are "surprised" and mistake the increase (decrease) as a change in the relative price of their own output, resulting in an increase (decrease) in the supply of output and employment in the economy. In the absence of such price surprises, output will remain at its natural level ($Y_{Nt}$).

Note that in Equation 4, a real variable ($Y$) is linked to a nominal variable ($P$). But, as Lucas demonstrates, the so-called classical dichotomy breaks down only when a change in the nominal variable is *unanticipated.*

Furthermore, Lucas (1996) notes that this distinction between anticipated and unanticipated monetary changes is a feature of all rational expectations-style models developed during the 1970s to explain the monetary nonneutrality exhibited in short-run trade-offs and, in his view, is one of the most important ideas in postwar macroeconomics.

## Theory, Policy Implications, and Empirical Evidence

During the 1970s, a burgeoning new classical research program led to a number of important and controversial contributions to macroeconomic analysis, in particular with respect to the Phillips curve, business cycles, the effectiveness and conduct of monetary and fiscal policy, and macroeconometric modeling and methodology.

### The New Classical Phillips Curve and Equilibrium Business Cycle Theory

In his seminal 1972 *Journal of Economic Theory* paper, Lucas demonstrated that within a Walrasian general equilibrium framework, monetary changes have real consequences, but "only because agents cannot discriminate perfectly between monetary and real demand shifts" so "there is no usable trade-off between inflation and real output" (p. 119). In Lucas's model, "the Phillips curve emerges not as an unexplained empirical fact, but as a central feature of the solution to a general equilibrium system" (p. 122). How can this be?

The answer provided by Lucas relates to agents (workers, households, firms) having imperfect information about their relative prices. As noted earlier, if agents are used to a world of price stability, they will tend to interpret an increase in the price of the good (or service) they produce as a relative price increase and supply more in response. Therefore, an unexpected increase in the price level will surprise agents, and they will misinterpret the information they observe. Agents have what Lucas (1977) refers to as a "signal extraction problem," and if all agents make the same error, the aggregate data will display a strong positive correlation between output and the general level of prices. Because Lucas's model is "monetarist," the increase in the general price level is caused by a prior increase in the money supply, and we therefore observe a positive money to output correlation (the nonneutrality of money).

Building on these key insights, the intellectual challenge facing Lucas was to provide a coherent and plausible monetary theory of aggregate instability in real variables (business cycles and the nonneutrality on money) in a world inhabited by rational profit-maximizing agents and where all markets continuously clear. His main innovation was to extend the classical model so as to allow agents to have *imperfect information.* As a result, Lucas's (1975) MEBCT has come to be popularly known as the *misperceptions theory.*

In the MEBCT, the supply of output at any given time ($Y_t$) has both a permanent (secular) component ($Y_{Nt}$) and a cyclical component ($Yc_t$), as shown in Equation 5:

$$Y_t = Y_{Nt} + Yc_t. \tag{5}$$

The permanent component of gross domestic product (GDP) reflects the underlying growth of the economy and follows the trend line given by Equation 6:

$$Y_{Nt} = \lambda + \Phi_t. \tag{6}$$

The cyclical component is dependent on the price surprise together with the previous period's deviation of output from its natural rate as shown in Equation 7:

$$Yc_t = \alpha\ [P_t - E(P_t \mid \Omega_{t-1})] + \beta\ (Y_{t-1} - Y_{Nt-1}). \tag{7}$$

The lagged output term in Equation 7 implies that deviations in output from the trend will be more than transitory due to the influence of a variety of propagation mechanisms, and the coefficient $\beta > 0$ determines the speed with which output returns to its natural rate after a shock. The observed serially correlated movements in output and employment (persistence) have been explained via a number of mechanisms that include reference to lagged output, investment accelerator effects, information lags and the durability of capital goods, and the existence of contracts inhibiting immediate adjustment and adjustment costs.

By combining Equations 5 to 7, we get the modified Lucas aggregate supply relationship given by Equation 8:

$$\begin{aligned} Y_t = {} & \lambda + \Phi_t + \theta\alpha\,[P_t - E(P_t / \Omega_{t-1})] \\ & + \beta(Y_{t-1} - Y_{Nt-1}) + \varepsilon_t, \end{aligned} \tag{8}$$

where $\varepsilon_t$ is a random error process, and $\theta$ represents the fraction of total individual price variance due to relative price variation. Thus, the larger is $\theta$, the more any observed variability in prices is attributed by economic agents to a real shock (i.e., a change in relative price) and the less it is attributed to purely inflationary (nominal) movements of the general price level. As Lucas's (1973) empirical paper demonstrates, this implies that monetary disturbances (random shocks) are likely to have a significant impact on real variables in countries where price stability has been the norm. In countries where agents are used to high inflation, monetary disturbances are unlikely to have any large impact on real variables. A major policy implication of the MEBCT is that a benign monetary policy would eliminate a large source of aggregate instability. Thus, new classical economists come down firmly on the side of rules in the "rules versus discretion" debate over the conduct of monetary policy.

## Policy Ineffectiveness Proposition

Although implicit in the early new classical papers by Lucas, the new classical policy ineffectiveness proposition was first formally presented in a highly controversial paper by Thomas Sargent and Neil Wallace (1975). The proposition can best be illustrated using the conventional aggregate demand/supply (AD/AS) model shown in Figure 39.1. Assume that an economy is initially operating at point A, where the price level ($P_A$) is fully anticipated and output and employment (and by implication, unemployment) are at their long-run (full information) equilibrium (natural) levels. Suppose the monetary authorities announce that they plan to increase the money supply from $M_1$ to $M_2$. Rational economic agents will take this information into account in forming their expectations and fully anticipate the effects of the increase in the money supply on the general price level, so that output and employment will remain unchanged at their natural levels. The rightward shift of the aggregate demand curve from $AD_1(M_1)$ to $AD_2(M_2)$, induced by an *announced* monetary expansion, will be completely offset by an upward shift of the short-run aggregate supply curve from $SRAS_1$ to $SRAS_2$, as money wages increase in response to an upward revision of price expectations. In the case of an anticipated monetary expansion, the economy will move immediately from point A to C, with no change in output and employment even in the short run (monetary neutrality). In contrast, if the monetary authorities surprise economic agents by engineering an unanticipated increase in the money supply (a monetary shock), rational economic agents with incomplete information will misperceive the resultant increase in the general price level as an increase in relative prices and react by increasing the supply of output and labor.

In terms of Figure 39.1, the aggregate demand curve would shift to the right from $AD_1(M_1)$ to $AD_2(M_2)$, to intersect $SRAS_1$ at point B. Output ($Y$) will deviate from its natural level ($Y_{Nt}$) as a consequence of deviations of the price level ($P_B$) from its expected level ($P_A$). Any increase or decrease in output or unemployment will only be temporary because once agents realize that there has been no change in relative prices, output and employment will return to their long-run equilibrium (natural) levels. In terms of Figure 39.1, as agents fully adjusted their price expectations the aggregate supply curve shifts upward, from $SRAS_1$ to $SRAS_2$ to intersect $AD_2$ at point C. The analysis is symmetrical with respect to a fall in aggregate demand to $AD_3(M_3)$. Here an *unexpected* negative monetary demand shock leads the economy to follow a path illustrated by points D and E, whereas an *expected* negative monetary demand shock will move the economy from A to E.

The policy ineffectiveness proposition has major implications for the controversy over the role and conduct of macroeconomic stabilization policy. If the money supply is determined by the authorities according to some "known" rule, then they will be unable to influence output and employment even in the short run. The use of a *systematic* monetary policy will be anticipated by agents. Hence, only departures from a known monetary rule will

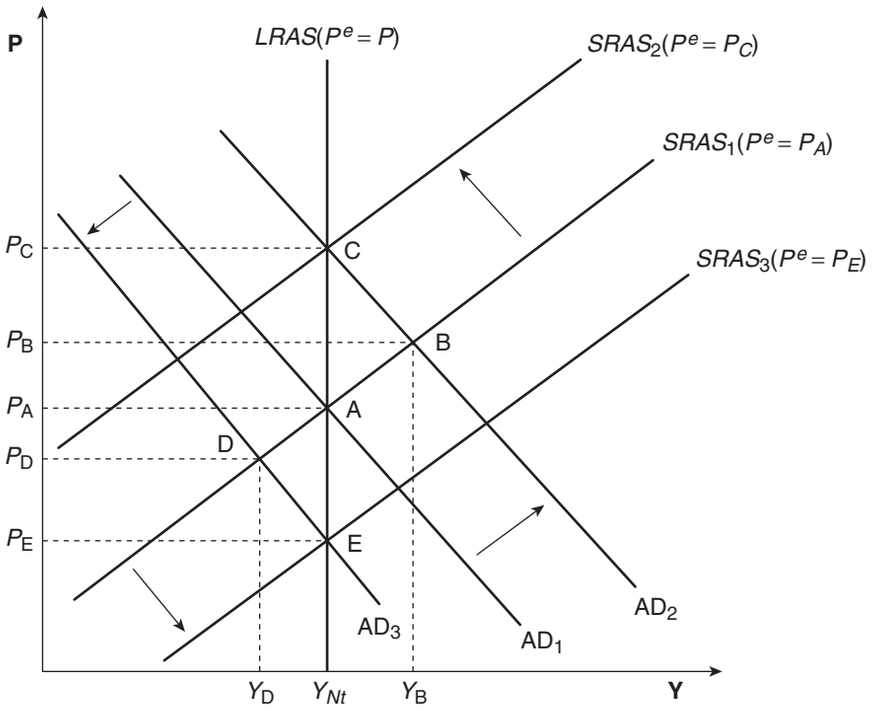


**Figure 39.1** Monetary Policy Effectiveness and Expectations

influence real variables. An important implication of this analysis is that a credible announced policy of disinflation will result in little or no sacrifice ratio in terms of output and employment because expectations of inflation will adjust instantaneously to the rate of inflation consistent with the announced reduced rate of monetary expansion. In this new classical scenario, only supply-side policies that influence the natural rates of output and unemployment will have lasting effects on real aggregate economic activity. However, as new Keynesian models demonstrate, monetary policy can be effective in a world of rational expectations once the assumption of continuous market clearing is abandoned.

## The Lucas Critique

The Keynesian macroeconometric models developed during the 1950s and 1960s by Nobel Laureates Lawrence Klein, Franco Modigliani, and others consisted of "systems of equations" involving endogenous variables and exogenous variables based on the well-known extended IS/LM–AD/AS framework. Such models contain *structural equations* comprising (a) identities, (b) institutional rules, (c) technological constraints, and (d) reduced form behavioral equations that describe the way in which individuals or groups will respond to the economic environment. These models were used for forecasting purposes and to test the likely impact of stochastic or random shocks. The model builders used historical data to estimate the model and then utilized the model to analyze the likely consequences of alternative policies.

In one of his most influential papers, "Econometric Policy Evaluation: A Critique," Lucas (1976) attacked the established Keynesian practice of using large-scale macroeconometric models to evaluate the consequences of alternative policy scenarios. Such policy simulations are based on the critical assumption that the estimated parameters of the model remain unchanged even when there is a change in the policy environment. Lucas argues that the parameters of large-scale macroeconometric models will not remain constant (invariant) in the face of policy changes, because rational economic agents will adjust their expectations and behavior to the new environment. The assumption of rational expectations implies that there will be interdependence between the prevailing policy framework and the parameters of the underlying behavioral equations. Macroeconometric models should thus take into account the fact that any change in policy will systematically alter the structure of the macroeconometric model. Because private sector structural behavioral relationships are noninvariant when the government policy changes, estimating the effect of a policy change requires knowing how economic agents expectations will change in response to the policy change.

This weakness of Keynesian-style macroeconometric models was particularly exposed during the 1970s as inflation accelerated and unemployment increased. The experiences of the 1950s and 1960s had led some policy makers and economic theorists to believe that there was a stable long-run trade-off between inflation and unemployment. However, once policy makers, influenced by this idea, shifted the policy regime and allowed unemployment to fall and inflation to rise, the Phillips curve shifted as the expectations of economic agents responded to the experience of higher inflation. Thus, by 1978, Lucas and Sargent famously declared that "existing Keynesian macroeconometric models are incapable of providing reliable guidance in formulating monetary, fiscal and other types of policy" (p. 69). To new classical economists, the predictions of orthodox Keynesian models had turned out to be "wildly incorrect" and a "spectacular failure," being based on a doctrine that was "fundamentally flawed" (p. 49).

The Lucas critique implies that the building of macroeconometric models needs to be wholly reconsidered so that the equations are truly structural or behavioral in nature—that is, independent of the policy regime. Ultimately, the influence of the Lucas critique contributed to the methodological approach adopted in the 1980s by modern new classical theorists of the business cycle, namely, *real business cycle* theories. Such models attempt to derive behavioral relationships within a dynamic optimization setting. According to Finn Kydland and Edward Prescott (1996), modern equilibrium models, "where empirical knowledge is organised around preferences and technologies" (p. 83), are free of the difficulties associated with Keynesian macroeconometric models and can account for the main quantitative features of business cycles.

## Dynamic Time Inconsistency, Credibility, and Monetary Policy Rules

The importance of credibility of pronouncements on monetary policy by central banks has long been appreciated. However, during the Keynesian era, economists and policy makers did not fully comprehend the link between policy outcomes and the credibility of the policy maker. This changed after the publication of the classic Kydland and Prescott (1977) contribution, "Rules Rather Than Discretion: The Inconsistency of Optimal Plans." This paper provided a clear exposition of *why* credibility is so important to the effective conduct of monetary policy. Consequently, it has had a profound influence on the long-standing rules versus discretion debate in macroeconomics, and on the conduct of monetary policy. Therefore, it was no surprise when Kydland and Prescott were awarded the 2004 Nobel Prize in Economics "for their contributions to dynamic macroeconomics: the time inconsistency of economic policy and the driving forces behind business cycles" (see Tabellini, 2005).

Kydland and Prescott provide a reformulation of the case against discretionary policies by developing an analytically rigorous new classical model where the monetary authorities are engaged in a strategic dynamic game with sophisticated forward-looking private sector agents. In this

setting, discretionary monetary policy leads to an equilibrium outcome involving too much inflation (an *inflation bias*). The difference between *ex ante* and *ex post* optimality is known as *time inconsistency.* Hence, what is an optimal policy announced at time $t$ will be time-inconsistent if a reassessment by the policy maker at $t + n$ implies a different optimal policy. Consequently, discretionary policies are incapable of achieving an optimal outcome combining low inflation at the natural rate of unemployment.

Kydland and Prescott assume that the monetary authorities can control the rate of inflation via monetary policy, that markets clear continuously, that economic agents have rational expectations, and that there is some social objective function ($S$), which rationalizes the policy choice and is of the form shown in Equation 9:

$$S = S(P^*_t, U_t), \text{ where}$$

$$S'(P^*_t) < 0, \text{ and } S'(U_t) < 0. \tag{9}$$

The social objective function (Equation 9) indicates that inflation and unemployment are "bads" because the negative first derivatives of this function, $S(P^*_t) < 0$ and $S(U_t) < 0$, imply that an increase (decrease) in inflation reduces (increases) social welfare and an increase (decrease) in unemployment also reduces (increases) social welfare. Equation 10, a variant of the rational expectations augmented Phillips curve, indicates that unemployment can be reduced by a positive inflation surprise:

$$U_t = U_{Nt} + \Phi\,(E(P^*_t \mid \Omega_{t-1}) - P^*_t). \tag{10}$$

Here, $U_t$ is unemployment in time period $t$, $U_{Nt}$ is the natural rate of unemployment, $\Phi$ is a positive constant, $E(P^*_t \mid \Omega_{t-1})$ is the rational expected rate of inflation, and $P^*_t$ the actual rate of inflation in time period $t$.

Suppose an economy is initially at a suboptimal but time-consistent equilibrium where actual and expected inflation equals 20% and unemployment is at the natural rate $U_{Nt}$. The policy maker aims to move the economy to the optimal position where actual and expected inflation equals zero at the natural rate of unemployment. To achieve this objective, the monetary authorities announce a reduction of the growth rate of the money supply to a rate consistent with zero inflation. If such an announcement is credible and believed by private economic agents, then they will lower their inflationary expectations, causing the Phillips curve to shift downward until zero inflation is achieved. But in this new situation, the monetary authorities may be tempted to renege on their promise to disinflate and engineer an inflationary surprise. If they exercise their discretionary powers and increase the rate of monetary growth in order to create an "inflation surprise," they can lower unemployment. However, this outcome is unsustainable, because unemployment is now below the natural rate and actual inflation exceeds expected inflation. Rational agents will soon realize they have been fooled, and the economy will return to the high inflation time-consistent equilibrium.

What this example illustrates is that the only way to achieve the optimal combination of inflation and unemployment is for the monetary authorities to establish credibility by abandoning discretion and precommitting to a noncontingent monetary rule consistent with price stability along the lines previously advocated by Milton Friedman. However, as became evident in the 1980s "monetarist experiment," the hard-core monetarist k-percent monetary growth rate rule proved to be unworkable on practical grounds due to instability in the velocities of monetary aggregates. An alternative approach to the achievement of price stability was required, and by the early 1990s, in many countries, this took the form of a two-pronged strategy to overcome the credibility problem in monetary policy, namely, (1) appointing an inflation-averse central banker (an *inflation hawk*) and (2) establishing an institutional structure specifying a strong nominal anchor (an explicit inflation target) combined with a guarantee of central bank independence.

## Ricardian Equivalence

While there were significant methodological differences between Friedman and Lucas, much of the early new classical literature was *monetarist* in orientation and concentrated on investigating the connection between monetary factors and aggregate instability. However, there are also significant implications for fiscal policy once agents are assumed to have rational expectations and are therefore forward-looking. In the standard Keynesian (IS/LM) model, a bond-financed increase in government expenditures will increase aggregate demand and, assuming the economy has spare capacity, will be expansionary on output and employment (government debt will have a positive wealth effect on consumption expenditures).

Robert Barro's (1974) seminal paper presents a modern exposition of David Ricardo's idea that government bonds should not be regarded as net wealth. The controversial Ricardian equivalence theorem (debt neutrality) states that the burden of government expenditure on the private sector will be identical whether it is financed by an increase in taxation or by the sale of government bonds. The latter involves a burden on the private sector that will take the form of future tax liabilities that are necessary to service the interest payments on, and redemption of, the bonds. According to new classical economists, rational, forward-looking private sector agents will take this future tax liability fully into account, and in consequence, government bonds will be not be regarded as net wealth. The present value of future tax liabilities will be expected to completely offset the value of the bonds sold. Because the permanent income hypothesis states that current consumption depends on the discounted present value of lifetime income, it does not matter if taxes are collected today or in the future. Therefore, the private sector will smooth their consumption

over time and react to a bond-financed increase in government expenditure by saving more (reducing consumption) in the present period in order to meet future tax liabilities.

## NCMII: Real Business Cycle Theory

The lack of clear supporting evidence from econometric work on the causal role of unanticipated monetary shocks in economic fluctuations was generally interpreted as providing a strong case for shifting the direction of new classical research toward models where real forces play a crucial role (Barro, 1978; Gordon, 1982; Laidler, 1986; Mishkin, 1982; Tobin, 1980). Consequently, in the early 1980s, Kydland and Prescott (1982) sought to provide a rigorous nonmonetary equilibrium account of the business cycle free from the theoretical flaws of earlier new classical models. The outcome has been the development and evolution of real equilibrium business cycle theory (REBCT) where the *impulse* mechanism of the earlier models (unanticipated monetary shocks) is replaced with supply-side (productivity) shocks in the form of random changes in technology. Real business cycle theorists also claim that their theories provide a better explanation of the "stylized facts" that characterize aggregate fluctuations.

In his 1977 paper "Understanding Business Cycles," Lucas advised that to understand business cycles, this could best be achieved "by constructing a *model* in the most literal sense: a fully articulated artificial economy which behaves through time so as to imitate closely the time series behaviour of actual economies" (p. 11). Such artificial economic systems can serve as laboratories "in which policies that would be prohibitively expensive to experiment with in actual economies can be tested out at much lower cost" (Lucas, 1980, p. 696). This methodological approach, via the use of computational experiments (calibration techniques), is exactly the approach adopted by real business cycle theorists during the 1980s (Kydland & Prescott, 1996). Lucas's (1977, 1980) papers can therefore be regarded as providing inspiration for the modern era of new classical equilibrium theorizing.

The REBCT research program starts from the position that aggregate instability in the form of business cycles and long-run growth are not separate issues. Following the methodology of Ragnar Frisch, real business cycle theorists distinguish between *impulse* and *propagation* mechanisms. An impulse mechanism is the initial shock that causes a variable to deviate from its steady state value. A propagation mechanism consists of those forces that carry the effects of the shock forward through time and cause the deviation from the steady state to persist. The major changes from MEBCT are with respect to (a) the dominant impulse factor, with technological shocks replacing monetary shocks; (b) the abandonment of the emphasis given to imperfect information as regards the general price level that played such a crucial role in the earlier monetary misperception models inspired by Lucas; and (c) the breaking down of the short-run/long-run dichotomy in macroeconomic analysis by integrating Robert Solow's neoclassical theory of growth with the theory of fluctuations (Plosser, 1989).

Although some versions of real business cycle theory allow for real demand shocks, such as changes in preferences or government expenditures, to act as the impulse mechanism, these models are more typically driven by exogenous productivity shocks. These stochastic fluctuations in factor productivity are predominantly viewed as the result of large random variations in the rate of technological change typically estimated using the "Solow residual" (Plosser, 1989). The conventional Solow neoclassical growth model postulates that the growth of output per worker over prolonged periods depends on technological progress that is assumed to take place *smoothly* over time. Real business cycle theorists reject this view and emphasize the *erratic* nature of technological change, which they regard as the major cause of changes in aggregate output. Controversially, it is also clear that, in order for real business cycle theories to explain the substantial variations in employment observed during aggregate fluctuations, there must be significant elasticity of labor supply via the *intertemporal substitution* of leisure. Furthermore, because in these models it is assumed that prices and wages are completely flexible, the labor market is always in equilibrium. In such a framework, workers *choose* unemployment or employment in accordance with their preferences and the opportunities that are available. To many economists, especially to those with a Keynesian orientation, this explanation of labor market phenomena remains unconvincing, and the claim by REBCT that aggregate fluctuations are the equilibrium outcomes of agents' efficient responses to productivity shocks is implausible (Mankiw, 1989). Moreover, REBCT cannot explain the initial causes of the 1930s Great Depression, even if it can offer a plausible if controversial account of why this traumatic event lasted so long (Pensieroso, 2007).

While recognizing these limitations of the REBCT approach, it is also important to recognize the positive contribution that this field of research has made in demonstrating that equilibrium models are not inconsistent with aggregate instability. By challenging theorists on all sides to recognize just how deficient our knowledge is of business cycle phenomena, the real business cycle approach has performed a useful function in raising profound questions relating to the cause, meaning, welfare significance, and characteristics of economic fluctuations (Rebelo, 2005).

## The Current Consensus and Future Directions

The transformation of macroeconomics during the last 40 years has been considerable, and the influence and contribution of economists such as Robert Lucas and Edward Prescott has been enormous. Their emphasis on deriving aggregate relationships that are based on firm

microeconomic foundations, and their interpretation and modeling of aggregate fluctuations as equilibrium phenomena, has revolutionized the way economists approach business cycle theory. By insisting on the importance of involving optimizing agents, the new classical economists have profoundly influenced the research agenda within macroeconomics (Prescott, 2006). However, the extent and channels of influence of new classical macroeconomics on policy also remain controversial, and in response to aggregate instability, new Keyneian economists continue to argue the case in favor of welfare-enhancing interventions via the use of monetary policy (Chari & Kehoe, 2006; Mankiw, 2006; Tobin, 1980).

While the defects in the Keynesian models of the neoclassical synthesis period have now been widely recognized and conceded by the vast majority of economists, it is also clear that they have been far from fatal. New Keynesian economists have responded to the new classical challenge by attempting to fix the theoretical problems raised by Lucas and have provided more coherent microfoundations for their macroeconomic models (Gordon, 1990; Snowdon & Vane, 2005). As a result, during the last decade of the twentieth century, the two main strands of macroeconomics began to converge into what Marvin Goodfriend and Robert King (1997) have labeled the *new neoclassical synthesis*. This new synthesis incorporates elements from both the new classical and new Keynesian perspectives into a coherent single framework. The central elements of this new synthesis comprise (a) the need for macroeconomic models to take into account intertemporal optimization; (b) a methodological preference for using dynamic stochastic general equilibrium models; (c) the widespread use of the REH; (d) recognition of the importance of imperfect competition in goods, labor, and credit markets; and (e) incorporating costly price adjustment and other imperfections into macroeconomic models (Woodford, 2008). More recently, some eminent economists, dissatisfied with the new synthesis, have advocated a more radical change of direction for macroeconomics that involves taking a behavioral approach to macroeconomic analysis (Akerlof & Shiller, 2009).

## References and Further Readings


Akerlof, G. A., & Shiller, R. (2009). *Animal spirits: How human psychology drives the economy and why it matters for global capitalism.* Princeton, NJ: Princeton University Press.

Barro, R. J. (1974). Are government bonds net wealth? *Journal of Political Economy, 82,* 1095–1117.

Barro, R. J. (1978). Unanticipated money, output and the price level in the United States. *Journal of Political Economy, 67,* 101–115.

Blanchard, O. J. (2000). What do we know about macroeconomics that Fisher and Wicksell did not? *Quarterly Journal of Economics, 115,* 1375–1409.

Chari, V. V., & Kehoe, P. J. (2006). Modern macroeconomics in practice: How theory is shaping policy. *Journal of Economic Perspectives, 20,* 3–28.

Fischer, S. (1996). Robert Lucas's Nobel memorial prize. *Scandinavian Journal of Economics, 98,* 11–31.

Friedman, M. (1968). The role of monetary policy. *American Economic Review*, *58*, 1–17.

Goodfriend, M., & King, R. (1997). The new neoclassical synthesis and the role of monetary policy. *National Bureau of Economic Research Macroeconomics Annual*, *12,* 231–283.

Gordon, R. J. (1982). Price inertia and policy ineffectiveness in the United States, 1890–1980. *Journal of Political Economy, 90,* 1087–1117.

Gordon, R. J. (1990). What is new-Keynesian economics? *Journal of Economic Literature, 28,* 1115–1171.

Hoover, K. D. (1988). *The new classical macroeconomics: A sceptical inquiry.* Oxford, UK: Basil Blackwell.

Keynes, J. M. (1936). *The general theory of employment, interest and money.* London: Macmillan.

Kydland, F. E., & Prescott, E. C. (1977). Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy, 85,* 473–491.

Kydland, F. E., & Prescott, E. C. (1982). Time to build and aggregate fluctuations. *Econometrica, 50,* 1345–1370.

Kydland, F. E., & Prescott, E. C. (1996). The computational experiment: An econometric tool. *Journal of Economic Perspectives, 10,* 69–85.

Laidler, D. E. W. (1986). The new classical contribution to macroeconomics. In B. Snowdon & H. R. Vane (Eds.), *A macroeconomics reader* (pp. 334–358). London: Routledge.

Lucas, R. E., Jr. (1972). Expectations and the neutrality of money. *Journal of Economic Theory, 4,* 103–124.

Lucas, R. E., Jr. (1973). Some international evidence on output–inflation tradeoffs. *American Economic Review, 63,* 326–334.

Lucas, R. E., Jr. (1975). An equilibrium model of the business cycle. *Journal of Political Economy, 83,* 1113–1144.

Lucas, R. E., Jr. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy, 1,* 19–46.

Lucas, R. E., Jr. (1977). Understanding business cycles. In K. Brunner & A. H. Meltzer (Eds.), *Stabilization of the domestic and international economy* (pp. 7–29). Amsterdam: North-Holland.

Lucas, R. E., Jr. (1980). Methods and problems in business cycle theory. *Journal of Money, Credit and Banking, 12,* 696–717.

Lucas, R. E., Jr. (1981). *Studies in business cycle theory.* Oxford, UK: Basil Blackwell.

Lucas, R. E., Jr. (1996). Nobel lecture: Monetary neutrality. *Journal of Political Economy, 104,* 661–682.

Lucas, R. E., Jr., & Sargent, T. J. (1978). After Keynesian macroeconomics. In *After the Phillips curve: Persistence of high inflation and high unemployment* (Federal Reserve Bank of Boston Conference Series No. 19, pp. 49–72). Boston: Federal Reserve Bank of Boston.

Mankiw, N. G. (1989). Real business cycles: A new Keynesian perspective. *Journal of Economic Perspectives, 3,* 79–90.

Mankiw, N. G. (2006). The macroeconomist as scientist and engineer. *Journal of Economic Perspectives, 20,* 29–46.

Mishkin, F. S. (1982). Does anticipated monetary policy matter? An econometric investigation. *Journal of Political Economy, 90,* 22–51.

Muth, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica, 29,* 313–335.

Pensieroso, L. (2007). Real business cycle models of the Great Depression: A critical survey. *Journal of Economic Surveys, 21,* 111–142.

Phelps, E. S. (1968, August). Money wage dynamics and labour market equilibrium. *Journal of Political Economy, 76,* 678–711.

Plosser, C. I. (1989). Understanding real business cycles. J*ournal of Economic Perspectives, 3,* 51–77.

Prescott, E. C. (2006). The transformation of macroeconomic policy and research: Nobel prize lecture. *Journal of Political Economy, 114,* 203–235.

Rebelo, S. (2005). Real business cycle models: Past, present and future. *Scandinavian Journal of Economics, 107,* 117–138.

Sargent, T. J., & Wallace, N. (1975). "Rational expectations," the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy, 83,* 241–254.

Snowdon, B., & Vane, H. R. (2005). *Modern macroeconomics: Its origins, development and current state.* Cheltenham, UK: Edward Elgar.

Tabellini, G. (2005). Finn Kydland and Edward Prescott's contribution to the theory of macroeconomic policy. *Scandinavian Journal of Economics, 107,* 203–216.

Tobin, J. (1980). Are new classical models plausible enough to guide policy? *Journal of Money, Credit and Banking, 12,* 788–799.

Woodford, M. (1999). *Revolution and evolution in twentieth-century macroeconomics.* Retrieved from http://www.columbia.edu/~mw2230

Woodford, M. (2008). *Convergence in macroeconomics.* Retrieved from http://www.columbia.edu/~mw2230

# PART V

## International Economics

# 40

# International Trade, Comparative and Absolute Advantage, and Trade Restrictions

Lindsay Oldenski
*Georgetown University*

Economists disagree about many aspects of economic policy. However, few topics garner as much support in the field as the potential for free trade to make individuals and societies better off. At the same time, protests of World Trade Organization meetings and labor union opposition to free trade agreements continue to make headlines. This chapter presents the theory underlying the conclusion that trade makes people better off and discusses conditions under which certain individuals or groups may not share in these gains from trade. Empirical tests of trade theory are also discussed.

This chapter is organized around the main theories, or models, of international trade. Each model seeks to explain certain aspects of the complex interactions between trading countries. Some models are more useful than others for answering certain questions. None tells the entire story, yet all capture important insights, and together they provide a useful basis for understanding international trade.

## Trade Theory and Evidence

### Ricardian Model of Comparative Advantage

*Absolute Advantage*

One simplistic view of world trade would be to expect that whatever country is "better" at producing a good in some absolute sense will end up specializing in the production of that good. Were this the case, it would spell bad news for poor developing countries considering opening up their borders to free trade. Because industrialized countries such as the United States have high levels of productivity across all sectors, a less technologically advanced developing country would have no hope of competing in a free trade environment if absolute productivity levels were all that mattered. For example, consider the United States and Mexico. Suppose that one laborer using U.S. technology can produce a computer in 2 hours of work. That same person working with U.S. agricultural technology can harvest a bushel of corn in 1 hour. Now suppose that in Mexico, producing a computer takes a person 12 hours, and harvesting a bushel of corn takes 3 hours. In this example, Mexico is slower at producing both computers and corn. We say that the United States has an *absolute advantage* in producing both goods because it can produce each of them at a lower cost (measured in person-hours) than Mexico.

*Comparative Advantage*

In 1817, a British economist named David Ricardo turned this idea of absolute advantage on its head. Using a model with two countries and two goods, he demonstrated that even if one country has an absolute advantage in the production of both goods, both countries can still gain from trade as long as their *relative* productivities for each good differ. The implications of this insight were huge. A country does not have to be highly developed or technologically advanced to reap the benefits of the global economy.

To see how this works, consider the example of the United States and Mexico described before. In this case, the United States is absolutely better at producing each good. However, relative productivities differ across the countries. In the United States, making one computer takes twice as long as harvesting a bushel of corn. So for each computer produced, the United States must forgo production of two bushels of corn. This tradeoff between the outputs of each good is known as the *opportunity cost* of production. The opportunity cost of producing a computer in the United States is two bushels of corn, and the opportunity cost of producing a bushel of corn in the United States is one half of a computer. However, in Mexico, the opportunity cost of producing a computer is much higher. Mexico must give up four bushels of corn for every computer produced. Yet the opportunity cost of producing a bushel of corn in Mexico is only one fourth of a computer. So while the United States has the absolute advantage in producing both goods, Mexico is relatively better at making corn (e.g., they do not have to give up as many computers for each unit of corn produced). We say that the United States has the comparative advantage in computers, and Mexico has the comparative advantage in corn.

### *The Pattern of Trade*

Both countries can benefit if they specialize based on comparative advantage. Sticking with this example, suppose that each country has 120 person hours available for production. This means that the United States can produce at most 60 computers or 120 bushels of corn. U.S. producers will most likely do something in between, say, dividing their labor evenly between the two sectors and producing 30 computers and 60 bushels of corn. If there is no possibility for international trade (a situation know as *autarky*), then they must consume exactly what they produce. Note that under autarky, in the United States, each additional computer the Americans want to consume requires them to give up 2 bushels of corn. Meanwhile, in Mexico, production is at most 10 computers or 40 bushels of corn and most likely something in between, such as 5 computers and 20 bushels of corn. In Mexico, each additional computer costs 4 bushels of corn. Now imagine that instead of making both goods, Mexico produces only corn and the United States produces only computers. When Mexico wants to give up some corn for computers or when the United States wants to give up some computers for corn, they can do so by trading on the world market. At what rate is the United States willing to give up its computers for corn? As long as they can get at least 2 bushels of corn for 1 computer, the United States is better off making only computers and then trading them for corn. And if Mexico can get at least one fourth of a computer for each bushel of corn (or only have to give up 4 bushels of corn or less for each computer), then they are better off producing only corn and trading for computers on the world market. The world relative price of corn and computers will end up being somewhere in between the opportunity costs in the two countries. For example, it could be 3 bushels of corn for 1 computer.

### *The Role of Wages*

In the Ricardian model, the pattern of trade is determined by relative labor productivity, not by wages. Even if wages were much lower in one country than in the other, the relative cost of producing each good within each country would still determine the gains from trade. It does not matter if the absolute cost of producing computers and corn is lower in a given country because they are more productive or because they pay lower wages. As long as there is some difference in the relative costs of producing each good across countries, then both countries can gain from trade.

### *Extensions to the Ricardian Model*

The standard formulation of the Ricardian model involves only two countries and two goods. However, the basic results still hold even with many countries or many goods. To incorporate many goods, simply line these goods up in order from the one in which the home country has the strongest comparative advantage (lowest opportunity cost) to the one in which the home country has the weakest comparative advantage (highest opportunity cost). There will be some cutoff point above which the home country will produce (and export) and below which the home country will import from its trading partner. So instead of completely specializing in one good, each country specializes in a subset of goods.

Incorporation of more than two countries can be handled in a similar way. When countries are lined up according to their labor productivity ratios in a given good, the country with the highest ratio will export that good and the country with the lowest ratio will import the good. Countries in the intermediate range may end up either exporting or importing the good, though all countries that export will have higher productivity ratios for that good than the countries that import.

### *Empirical Evidence*

The Ricardian model's simplicity, as well as its stark predictions, makes it difficult to test empirically. In the real world, production requires more inputs than just labor and we generally observe partial rather than complete specialization. However, the basic prediction of the model—that countries tend to export goods for which their productivity is relatively high—has strong empirical support. A classic 1963 study by Bela Balassa compared British and American labor productivity and trade. In the period studied, the United States had higher absolute labor productivity than the United Kingdom in almost all industries. Yet

British exports were equally as large as those of the United States. Balassa found that the goods being exported by Britain were those in which the country had a *relative* productivity advantage, even though absolute U.S. labor productivity was higher. More recent evidence shows that, while the United States had higher overall labor productivity than Japan in the 1990s, Japan's relative labor productivity in the automobile industry was about 20% higher than that of United States, potentially explaining the large volume of automobile exports from Japan to the United States (Krugman & Obstfeld, 2005)

## Specific Factors (Ricardo-Viner) Model

One key assumption in the Ricardian model is that people working in, say, an automobile manufacturing plant can immediately switch to a job programming software or giving management consulting advice should they lose their old job to the forces of free trade. Another way to say this is that the Ricardian model assumes that labor may move freely between sectors. The Ricardo-Viner model relaxes this assumption by allowing certain factors of production to be specific to (or used exclusively in) certain industries. In the classic version of the Ricardo-Viner model, land is specific to agricultural production, capital is specific to manufacturing, and labor can move freely between the two sectors. However, the implications would be the same if we assumed that, say, certain types of labor were specific to each sector and capital could move freely between them. Capital, land, and labor are all known as factors of production. What matters in this model is that three factors exist, one of them is mobile across sectors, and the other two can be used in only one specific sector.

The existence of these specific factors leads to more subtle outcomes than the stark predictions of complete specialization in the Ricardian model. Instead of trading off production of two goods at a constant rate (e.g., two bushels of corn for one computer), production in each sector exhibits diminishing returns. Consider the agricultural sector. The country is endowed with a fixed amount of land, and this land can be used only for agricultural production. The land on its own cannot produce output, however. Labor is also needed for agricultural production. In this example, labor is the factor that can move freely between the agricultural and manufacturing sectors. With no labor, production per acre of land is zero. When the first unit of labor moves to the agricultural sector, the marginal product of that one unit of labor is very large as production goes from zero to some positive amount. Each additional unit of labor also increases production, but not by nearly as much as that first unit. As more and more workers move from manufacturing to agriculture, these workers add less and less additional output (or marginal product) because they have only a fixed amount of land to work with. The same is true in the manufacturing sector. The first worker who shows up and turns on the machinery (or capital) has made an enormous contribution to output, yet the individual (or marginal) contribution of each additional worker is less and less given the fixed amount of capital. It follows that the optimal allocation of labor between the two sectors will involve at least some agricultural and some manufacturing production rather than complete specialization in one sector. It would not make sense for an economy to allow its last worker to have a very small impact working in a crowded factory when that worker could have an enormous productivity impact by moving to an empty field and producing agricultural goods. Wages are the adjustment mechanism that ensures this balance. Employers will hire an additional unit of labor up to the point at which the additional value produced by that unit of labor exactly equals the cost, or wage rate, of that labor. As described previously, the marginal product added by each additional worker in a sector is declining, so wages in a sector must also be declining as more and more workers move to that sector. Because labor can move freely across sectors, if workers could earn higher wages working in a factory than on a farm, they would quit their jobs in the farming sector and move into manufacturing instead (and vice versa). Equilibrium requires that the wage rates, and thus marginal products of labor, must be equal across sectors. Due to the diminishing product of labor, this is not likely to occur under complete specialization.

The relative price of the two goods in a country reflects the relative cost of producing those goods. If a factor of production is relatively scarce, then it will be relatively costly, translating into a relatively higher price for the good that uses the scarce factor in its production. On the other hand, the industry with the relatively abundant specific factor will have a relatively lower price. When two countries move from autarky to free trade, they are no longer constrained by their own factor endowments. For example, if one country is relatively well endowed with land versus capital when compared to another country, then agricultural goods will be relatively cheaper and manufactured goods will be relative more expensive in the first country. When the two countries open up to trade, Country 1 now has the option of importing some cheaper manufactured goods from abroad rather than making them itself. This reduces that demand for the scarce capital in Country 1, reducing both the cost of capital and the price of manufactured goods in that country. A similar process occurs with the relatively scarce land in Country 2. Trade allows this country to import some agricultural goods from Country 1, freeing up land and reducing the price of domestically produced food. At the same time, the demand for Country 1's abundant land and cheap agricultural products goes up, raising their price, while the relative price of Country 2's abundant capital and manufacturing products also goes up. This happens until relative prices are equalized across the two countries. The end result is that the country that is relatively abundant in land sees the relative price of goods that use land go up

and the relative price of goods that use capital go down. The opposite occurs in the country that is relatively abundant in capital.

As in the Ricardian model, the country as a whole is better off under free trade than under autarky. Each country sees the relative price of its exported good go up, which increases income from exports. At the same time, the price of the imported good goes down, meaning that they can buy more of the imported good for a given amount of export production.

However, the existence of specific factors leads some individuals to be hurt by trade. Owners of the factor specific to the export industry gain from trade as the demand for their factor of production increases. Owners of the factor specific to the import-competing sector are worse off because the demand for (and thus the returns to) their factor have gone down. In the example just described, owners of the abundant land in Country 1 see the demand for their land go up and thus receive higher rental income for that land. However, the owners of the scarce capital see the demand for their factor go down and thus receive a lower return to that capital. The impact on the mobile factor (in this case, labor) is ambiguous, because workers will have to pay more for the exported good but can now consume the imported good more cheaply. This theoretical result can explain why certain individuals or interest groups may be opposed to free trade even if the world as a whole gains from trade.

### *Empirical Evidence*

In addition to partial specialization along the lines of comparative advantage, this model also predicts that (a) each country will have both winners and losers from trade and (b) these winners and losers can be distinguished based on the industries that they are connected with. Precisely quantifying the gains and losses from trade that accrue to specific groups is not something that can be easily done with existing data. However, if we assume that individuals will either be for or against more open trade policies based on the perceived impact on their own welfare, then we can use the organization of interest groups and their voting patterns on trade policy proposals to test these predictions. If this model is correct, then we should observe support for and opposition to trade-related legislation to be divided by industry. Stephen Magee (1980) finds that this is indeed the case.

## Heckscher-Ohlin Model

In the Heckscher-Ohlin model, comparative advantage comes from an interaction between the characteristics of countries and industries. It is a model with two countries, two goods, and two factors of production. Each country is endowed with a certain amount of each factor (e.g., high-skilled and low-skilled labor). By comparing the ratio of high-skilled to low-skilled labor in each country, we can say that the country with the larger ratio is relatively well endowed with (or abundant in) high-skilled labor and the country with the smaller ratio is relatively well endowed with (or abundant in) low-skilled labor. A similar comparison can be made between goods by looking at the ratio of high- to low-skilled labor used in the production of each good. The good with the higher ratio is said to be intensive in the use of high-skilled labor, and the good with the lower ratio is intensive in the use of low-skilled labor. The Heckscher-Ohlin model can be summarized by its four main theorems.

1. *The Heckscher-Ohlin Theorem.* The country that is relatively more abundant in a factor of production will export the good that uses that factor relatively intensively. In other words, countries trade based on comparative advantage, and the source of that comparative advantage is relative factor endowments interacted with relative factor intensities.

2. *The Rybczynski Theorem.* When the endowment of a factor increases, output of the good that uses that factor intensively increases more than proportionally and output of the other good falls.

This theorem links relative factor endowments to production. Even under autarky, this relationship holds. It is not surprising that an increase in one factor should increase production of the good that uses that factor intensively. However, the coinciding reduction in output of the other good is a unique result of this model.

3. *The Factor Price Equalization Theorem.* If goods prices are equalized by trade, then so are factor prices.

This result relies on several key assumptions. The first is that of perfect competition. Under perfect competition, goods are priced at their marginal cost of production. If this were not the case, another producer would be able to enter the market and sell the identical good at a lower price. Marginal cost pricing results in a direct relationship between goods prices and factor prices. Another crucial assumption is that production technology is identical across countries, such that one unit of capital or labor will have the same marginal product in each country. Otherwise, we could observe marginal cost pricing and goods prices that are equalized across countries, even if wages are much lower in one country, if the low wage country is less productive. For example, if one country requires 10 person hours to produce a good and the wage is $1.00 per hour, then the cost per unit of good is the same as in a country where production requires only 1 person hour but the wage is $10.00 per hour. This result also assumes that wages are set by market forces rather than trade unions or government policies.

4. *The Stolper-Samuelson Theorem.* When relative prices change, the factor used intensively by the good whose

price has increased receives a greater reward in terms of both goods.

Trade creates clear winners and losers within each country. The winners are those who control the relatively abundant factor, and the losers are those who control the relatively scarce factor. Suppose a country is relatively well endowed with high-skilled labor. When that country opens up to trade with a relatively low-skilled country, they will specialize in high-skill intensive goods. This increases the demand for high-skilled labor and raises the wages of high-skilled workers. At the same time, there is less demand for low-skilled workers, because the goods they produce can be purchased more cheaply elsewhere. Low-skilled workers in the high-skill abundant country will see their wages fall. Focusing on high- and low-skilled workers as the two factors of production also has implications for the impact of trade on the income distribution within countries. Because more developed countries are more well endowed with high-skilled labor, trade should cause them to specialize in high-skill intensive goods, increasing the wages of high-skilled workers and decreasing the wages of low-skilled workers. Because high-skilled jobs generally pay more, this results in a wider wage gap between high-income and low-income workers. However, the opposite prediction holds for developing countries. When these countries, which are relatively more well endowed with low-skilled labor, open to trade the wages of low-skilled workers should increase, leading to a reduction in income inequality.

### *Empirical Evidence*

John Romalis (2004) tests a version of the Heckscher-Ohlin model with many countries and many goods. He finds strong empirical evidence for the prediction that countries capture a larger share of world trade in sectors that use their abundant factors more intensively. He also finds evidence in support of the Rybczynski theorem. James Harrigan (1995) also finds support for the Rybczynski theorem by looking at the impact of changes in capital, skilled labor, and unskilled labor on industry-level production using a panel of 20 countries over 15 years.

The factor price equalization theorem is less well supported. A literal interpretation of this theorem would predict that wages and rental rates should be equal across all countries. This is clearly not the case. However, Paul Samuelson (1971) suggested that factor price convergence might be a better concept to test for than factor price equalization, because wages have been getting closer across countries as trade has increased even though they are clearly not equal. One of the strongest assumptions of the Heckscher-Ohlin model is that technology does not vary across countries. Edward Leamer and James Levinsohn (1995) summarize a number of empirical studies that suggest that once productivity differences have been accounted for, factor price equalization actually does have quite a bit of empirical support.

Support for the Stolper-Samuelson theorem is mixed. As mentioned previously, a study by Magee (1980) finds that support for trade policies is generally based on industry affiliation rather than factor affiliation. However, Ronald Rogowski (1987) does find support for the Stolper-Samuelson predictions in the formation of political coalitions. Adrian Wood (1997) examines the prediction that trade should increase the wage gap in developed countries and decrease the wage gap in developing countries. He finds strong evidence that trade is associated with increasing income inequality in developed countries, but the impact on developing countries is mixed. When several East Asian countries first opened to trade in the 1960s and 1970s, they experienced a narrowing of the wage gap. However, most Latin American countries that liberalized their trade policies in the 1980s saw their wage gaps widen. However, a number of potential mitigating factors, such as the role of skill-biased technological changes and the emergence of even more low-skill endowed countries such as China and India on the world market, make it difficult to sort out the extent to which the evidence supports or contradicts the Heckscher-Ohlin theory.

## Standard Trade Model

The standard trade model is not a separate model, but instead provides a general framework in which to view the other classical models. In this framework, comparative advantage can come from (a) productivity differences across countries as in the Ricardian model, (b) differences in endowments of industry-specific factors of production as in the specific factors model, or (c) interactions between the relative factor endowments of countries and relative factor intensities of goods as in the Heckscher-Ohlin model. By keeping the sources of comparative advantage in the background, this model is able to focus on the welfare effects of trade.

In the standard trade model, each country produces the greatest value of output that it can, given the prevailing relative prices. This determines the global relative supply curve. Each country consumes the bundle that will give it the highest utility it can afford at prevailing relative prices, which determines the global relative demand curve. World equilibrium is determined by the intersection of the relative supply and relative demand curves. The primary channel through which trade impacts welfare is through a country's *terms of trade,* which is defined as the price of the country's exports divided by the price of its imports. An increase in a country's terms of trade leads to an increase in welfare, as it allows the country to purchase more imports for the same quantity of exported goods. This framework can be used to examine how increased trade volumes, changes in preferences, trade policy instruments such as tariffs, growth, and income transfers impact relative world prices and thus the welfare of nations.

### Imperfect Competition Model

The classic trade models described previously emphasize ways in which differences between countries can lead to gains from trade. Yet most trade is between countries that are very similar to each other in terms of both technology and endowments. Paul Krugman developed a model in which imperfect competition can lead to gains from trade, even between identical countries. Economies of scale lead to imperfect competition. Under economies of scale, production is more efficient at a larger scale, possibly because there exist some large fixed costs to market entry. Consider an industry that requires a large amount of research and development, such as the pharmaceutical industry. To produce a drug, a firm in this industry must invest hundreds of millions of dollars and many years of research time before the product hits the market. Once these fixed costs have been incurred, however, the marginal cost of each additional dose of the drug produced is close to zero. If the company were to charge a price equal to the marginal cost of production, it would never recover its initial investment. For this reason, we expect prices to equal average cost rather than marginal cost. If the company produced only a few doses, the average cost of production per dose would be extremely high. Yet as it produces more and more doses, the average cost per dose goes down, as does the price paid by the consumer.

Trade increases the size of the market, leading to lower average costs of production. Under autarky, each type of medication would have to be produced by each country. Thus, the fixed costs would be incurred by more than one producer and the average costs would reflect the market size in each country. If two countries were able to trade, then they could specialize such that the high fixed costs of each product had to be incurred only once, and the larger world market size would lead to lower average costs and thus lower prices for consumers.

In his model of imperfect competition and international trade, Paul Krugman focuses on a type of imperfect competition known as *monopolistic competition.* In this framework, each firm produces a product that is differentiated from those of its rivals. Even if there are many firms in an industry, each firm is a monopolist in its own variety of the product. Some substitution between products is possible, but they are not perfect substitutes. Thus, each individual producer behaves as a monopolist even though there is some competition between firms. Consumers not only benefit from the lower prices that result from the concentration of production, but are also able to choose between a greater number of varieties in each industry. This framework allows for trade within industries as well as between them. Consider the automobile industry. The United States is both an exporter of Fords and an importer of Toyotas. Krugman's model can explain how countries that are similar in their technological development, such as the United States and Japan, can still gain from trade and can both be exporters of automobiles.

*Empirical Evidence*

The large share of intraindustry trade in total world trade flows is often cited as evidence of the monopolistic competition model. About 25% of world trade occurs within industries. Krugman points out that industries with higher levels of intraindustry trade tend to be sophisticated manufactured goods, such as chemicals, pharmaceuticals, and electrical machinery, which are likely to exhibit increasing returns to scale and allow for product differentiation. Goods with lower shares of intraindustry trade tend to be more labor-intensive products such as footwear and apparel (Krugman & Obstfeld, 2005). Elhanan Helpman (1987) tests the predictions of the monopolistic competition model using Organisation for Economic Co-operation and Development data for 1956 to 1981. He finds support for the prediction that the share of intraindustry trade should be higher for pairs of countries with similar per capita gross domestic product (GDP).

## Current Topics in International Trade Theory

### The Gravity Model

The gravity model, in which the volume of trade between two countries is based on the distance between them and their relative GDPs, as well as a number of other country characteristics such as population, language, common borders, colonial history, tariff rates, and so on, has been shown to be empirically very robust. While the initial motivation for the model was empirical, gravity has since been shown to be derivable from a variety of theoretical frameworks. Helpman and Krugman (1985) derive gravity predictions using a monopolistic competition model, Alan Deardorff (1998) provides a derivation using the Heckscher-Ohlin model, and Jonathan Eaton and Samuel Kortum (2002) provide a Ricardian treatment. For a survey of these and other derivations, see Robert Feenstra (2002).

### Heterogeneous Firms

Most classic trade models are based on the notion that countries trade in industries. However, in actual commerce, it is firms rather than countries that conduct the transactions. Within a given industry, there may be a number of different firms, each of which may have different levels of productivity or ways of organizing production. As more detailed data on firm level transactions have become available, and as increased computing power has allowed for the analysis of these detailed microdata, the focus of both empirical and theoretical research has moved toward including firm-level heterogeneity. Andrew Bernard and J. Bradford Jensen (1999) summarize evidence documenting that exporting firms have higher levels of productivity than nonexporters. They demonstrate that this is because the better firms select into exporting, not because exporting leads to higher

productivity. Marc Melitz (2003) develops a model based on Krugman's monopolistic competition framework that incorporates firm-level productivity differences and can explain key patterns in the firm-level trade data. Much of the current research on international trade incorporates some form of the Melitz-type model of heterogeneous firms.

## Trade Restrictions

### Tariffs

A *tariff* is a tax imposed on goods or services when they are sold across international borders. Ad valorem tariffs are a fixed proportion of the product's value (e.g., 10%). Specific tariffs are a fixed amount per unit (e.g., $10,000 per automobile) and do not vary with price. Tariffs harm consumers, who are forced to pay higher prices for the goods they consume. Tariffs also harm producers whose goods are being taxed. However, they benefit domestic producers who are in competition with the foreign goods that are subject to the tariff. They also create government revenue. These gains and losses are depicted in Figure 40.1. Even though some groups benefit and some groups lose when tariffs are imposed, the distortionary effect on prices leads to an efficiency loss such that the negative effects of the tariff outweigh the benefits that accrue to certain individuals.

### Nontariff Barriers

In addition to tariffs, which directly impact the prices paid by consumers, governments may also impose a variety of quantity-based restrictions on imports. An extreme example is an import ban, which prohibits any imports from entering the country. Import quotas, which limit the quantity of a certain good that may enter the country, are more common. Although they focus on quantities rather than prices, quota restrictions have the same effect as tariffs. Restricting the supply of a certain good will result in a price increase similar to the increase that could be observed under tariff policy. Other restrictions such as product standards or labeling requirements have a similar effect.

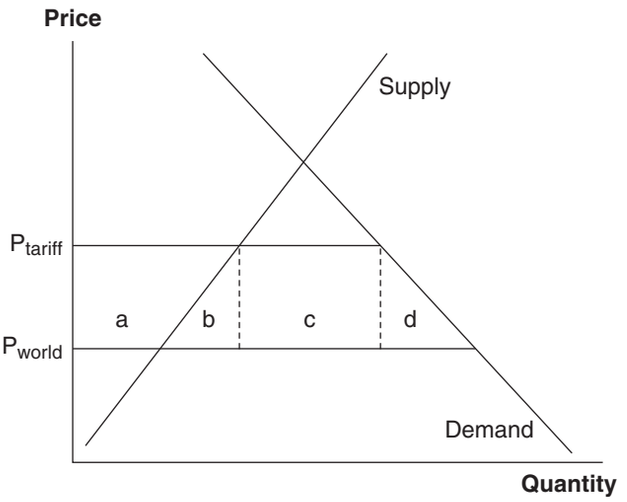


**Figure 40.1** Costs and Benefits of a Tariff

*Notes*: The imposition of a tariff raises the price that consumers pay from $P_{world}$ to $P_{tariff}$. The loss of consumer welfare due to this increase in price is equal to the area $a + b + c + d$. Area $a$ represents the gain to domestic producers, who may now charge a higher price for the goods they produce. Government revenue from the tariff is represented by area $c$. Areas $b$ and $d$ are pure efficiency loss.

### Preferential Trade Agreements

Preferential trade agreements (PTAs) are agreements in which the tariffs that participants apply to other members' goods are lower than the rates on the same goods coming from countries outside of the agreement. The two main types of PTAs are free trade areas and customs unions. In a *free trade area* (FTA), each country's goods can be shipped to the others without tariffs, but each country sets its own tariffs against nonmembers independently (e.g., North American Free Trade Agreement). A *customs union* is like a FTA, but the member countries agree on uniform tariff rates to apply to nonmembers (e.g., the European Union). FTAs are politically more straightforward because they do not require that member countries come to agreement on their external tariff policy. However, FTAs create a larger administrative burden than customs unions because they must prevent third-party imports from first entering the member country with the lowest external tariff and then being shipped to the member with the higher external tariff in order to avoid paying that tariff. *Trade diversion* is a potential negative consequence of PTAs. It occurs when a formal agreement prevents potentially beneficial trade with countries outside of the agreement.

## Policy Implications

Government policies other than tariffs and trade restrictions can also distort the predictions of the standard trade models. Differences in the cost of production across countries may not accurately reflect productivity or factor abundance in the presence of such distortions. For example, wages will not be fully determined by the demand for labor in countries with a strong labor union presence. Differences in environmental and workers' rights regulations may also impact production costs. However, the predictions of the models described in this chapter will still hold as long as the productivity and endowment differences across countries outweigh the impact of labor and environmental regulations.

The theoretical and empirical evidence points overwhelming in the direction of gains from trade. The natural policy conclusion is that restrictions on free trade are harmful and should be avoided. In spite of the net gains resulting from increased trade, there is also evidence that

these benefits do not accrue universally, and that trade creates both winners and losers. However, because the gains outweigh the losses, a standard argument made by trade economists is to point out that everyone can be made better off if countries liberalize their trade policies and then use some of the income gains resulting from this liberalization to compensate those that otherwise would have been made worse off by trade. For example, a lump sum tax could be imposed and the revenues from that tax could be used to provide an income subsidy to those who would have been made worse off.

## Conclusion

A number of theoretical models exist to explain the causes and effects of international trade patterns. The Ricardian model of comparative advantage was the first to demonstrate that any country can gain from trade, regardless of absolute levels of productivity. In this model, all that matters is that countries differ in their relative abilities to produce different goods. The basic result of the Ricardian framework, that specialization according to comparative advantage leads to gains from trade, has held up over time. However, the simplistic nature of the model means that it is not very well suited for answering more subtle questions, such as how trade impacts the distribution of income within a country or why we do not observe complete specialization. The specific factors (Ricardo-Viner) and Heckscher-Ohlin models add more depth to the comparative advantage story. In both of these models, countries only partially specialize. While each country as a whole gains from trade, individuals in import-competing industries are made worse off by trade, while those in exporting sectors are made better off. The Krugman model of trade under imperfect competition shows how even two identical countries can gain from trade in the presence of economies of scale and product differentiation. Current research on international trade focuses on making more direct links between the theory and data and emphasizes the role of firms, rather than countries and industries, in trade.

## References and Further Readings


Appleyard, D., Field, A., & Cobb, S. (2006). *International economics* (5th ed.). Boston: McGraw-Hill Irwin.

Balassa, B. (1963). An empirical demonstration of classical comparative cost theory. *Review of Economics and Statistics, 4,* 231–238.

Bernard, A., & Jensen, J. B. (1999). Exceptional exporter performance: Cause, effect, or both? *Journal of International Economics, 47*(1), 1–25.

Bhagwati, J., Panagariya, A., & Srinivasan, T. N. (1998). *Lectures on international trade* (2nd ed.). Cambridge: MIT Press.

Corden, W. M. (1997). *Trade policy and economic welfare* (2nd ed.). Oxford, UK: Oxford University Press.

Deardorff, A. (1998). Determinants of bilateral trade: Does gravity work in a neoclassical world? In J. Frankel (Ed.), *The regionalization of economy* (pp. 7–22). Chicago: University of Chicago Press.

Eaton, J., & Kortum, S. (2002). Technology, geography, and trade. *Econometrica, 70*(5), 1741–1779.

Feenstra, R. (2002). The gravity equation in international economics: Theory and evidence. *Scottish Journal of Political Economy, 49*(5), 491–506.

Feenstra, R., & Taylor, A. (2008). *International economics*. New York: Worth.

Harrigan, J. (1995). The volume of trade in differentiated intermediate goods: Theory and evidence. *Review of Economics and Statistics, 77*(2), 283–293.

Helpman, E. (1987). Imperfect competition and international trade: Evidence from fourteen industrialized countries. *Journal of the Japanese and International Economies, 1,* 62–81.

Helpman, E., & Krugman, P. (1985). *Market structure and foreign trade: Increasing returns, imperfect competition and the international economy.* Cambridge: MIT Press.

Krugman, P., & Obstfeld, M. (2005). *International economics theory and policy* (7th ed.). Boston: Pearson Addison-Wesley.

Leamer, E., & Levinsohn, J. (1995). International trade theory: The evidence. In G. Grossman & K. Rogoff (Eds.), *Handbook of international economics* (pp. 1339–1394). Amsterdam: Elsevier.

Magee, S. (1980). Three simple tests of the Stolper-Samuelson theorem. In P. Oppenheimer (Ed.), *Issues in international economics* (pp. 138–151). London: Oriel Press.

McKinsey Global Institute. (1993). *Manufacturing productivity.* Washington, DC: Author.

Melitz, M. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica, 71*(6), 1695–1725.

Rogowski, R. (1987). Political cleavages and changing exposure to trade. *American Political Science Review, 81,* 1121–1137.

Romalis, J. (2004). Factor proportions and the structure of commodity trade. *The American Economic Review, 94*(1), 67–97.

Samuelson, P. (1971). An exact Hume-Ricardo-Marshall model of international trade. *Journal of International Economics, 1,* 1–18.

Wood, A. (1997). Openness and wage inequality in developing countries: The Latin American challenge to East Asian conventional wisdom. *The World Bank Economic Review, 11*(1), 33–57.

# 41

# Balance of Trade and Payments

Lawrence D. Gwinn
*Wittenberg University*

The balance of payments registers the international financial position of a country, using a double-entry bookkeeping approach to tabulate the market value of the transactions in goods, services, and financial assets between the country's residents and the residents of the rest of the world. Like gross domestic product (GDP), the balance of payments encompasses transactions in goods and services produced during the year, but, unlike the GDP, the balance also encompasses transactions in assets. In addition to categorizing international transactions as debits or credits, the balance of payments separates private transactions in goods and services into the current account and transactions in assets into the capital account. Official government transactions, undertaken to affect the exchange rate, are typically separated from private transactions in balance of payments accounting.

While nations had been aware of the relationship between their exports and imports since they first coalesced into discernable entities and began to trade, in *History of Economic Analysis* (1954), Joseph Schumpeter likens the conceptualization of the balance of payments to an idea that has been vaguely known for centuries without being fully understood. Schumpeter credits Antonio Serra with having the first "clear conception" of the balance of payments in 1613. While Serra recognized and understood many of the mercantilist ideas, according to Schumpeter, his "most important point" was an understanding of the connection between a country's balance of payments position and its domestic economic situation. Schumpeter also mentions Thomas Mun (1571–1641) and other mercantilists as having an understanding of the balance of payments, notwithstanding their misapplication of its principles to trade policy.

## Classifications Used in Balance of Payments Accounting

The balance of payments categorizes international transactions along several interrelated lines. First, any transaction on the balance of payments is either a credit or a debit. A transaction that involves a payment received from foreign residents, due to the international sales of goods, services, or assets by domestic residents, is a credit on a country's balance of payments; a transaction that involves a payment to foreign residents is a debit. Second, the balance of payments classifies international transactions as either autonomous or accommodating. An autonomous transaction is undertaken independently of any other international transaction and gives rise to an accommodating transaction. Finally, the balance of payments accounts decompose international transactions into current account transactions, which represent trade in *currently produced* goods and services and capital account transactions, which represent trade in assets.

### Credits and Debits

As a result of the double-entry bookkeeping approach, every transaction appears on the balance of payments as either a credit or a debit. A transaction that involves a payment received from foreign residents, due to the international sales of goods, services, or assets by domestic residents, is a credit on a country's the balance of payments. For example, when a U.S. firm exports a product and receives a payment from a foreign resident, the value of the product appears on the U.S. balance of payments as a credit, with a positive sign, on the balance of payments

account. Alternatively, if foreign residents purchase U.S. government bonds from a U.S. bond dealer, the bond dealer receives a payment for the bonds, and this transaction also appears on the balance of payments as a credit.

The balance of payments classifies any transaction involving a payment to foreign residents as a debit. If a U.S. resident buys a foreign automobile, for example, making payment to the foreign auto firm, the payment is a debit and appears on the balance of payments accounts with a negative sign.

Moreover, as a result of the double-entry bookkeeping approach, every transaction appears twice on the balance of payments as both a credit and a corresponding debit, so that for every international transaction, there are two entries on the balance of payments accounts of equal magnitude, but opposite sign. For example, the purchase of a foreign automobile by a U.S. resident involves a payment to foreign residents and is therefore classified as a debit on the balance of payments. But the payment made to foreign residents for the car must always result in increased foreign holdings of U.S. assets, frequently a deposit in a U.S. commercial bank account, which the balance of payments records as a credit. Note that this approach is simply a conventional method of taking international transactions into account; there is nothing inherently *good* about a credit and nothing inherently *bad* about a debit.

## Autonomous and Accommodating Transactions

In addition to classifying international transactions as credits and debits, the balance of payments also categorizes international transactions as either autonomous or accommodating. For example, the purchase of a foreign automobile by a U.S. resident is a debit on the balance of payments and is autonomous because it is undertaken independently of any other international transaction. But because the payment made to foreign residents for the auto must always initiate an increase in foreign holdings of U.S. assets, such as the bank deposit, balance of payments accounting methods classify it as an accommodating transaction. Accommodating transactions are so named because they accommodate the autonomous transactions or allow the autonomous transaction to occur. The purchase of the automobile could not take place were it not accommodated by the increased holdings of U.S. assets by foreigners.

Conceptually, the best measure of the overall balance of payments is the sum of all autonomous credits and debits. An excess of autonomous credits over autonomous debits would indicate a balance of payments surplus, and an excess of autonomous debits over credits would indicate a balance of payments deficit. Thus, it would be ideal to distinguish all international transactions as described previously, separating them into their autonomous and accommodating categories. However, as we will see in the discussion of sample transactions, it is not possible to determine whether a given transaction is autonomous or accommodating, which means there is no satisfactory overall measure of the balance of payments.

## Current Account and Capital Account Balances

Finally, the balance of payments accounts decompose international transactions into current account transactions and capital account transactions. The current account mainly consists of trade in *currently produced* goods and services that are produced during the current year. All imports and exports as well as unilateral transfers, such as international gifts, go in the current account. Investment income goes in the current account, as well, because it represents a payment for the service currently provided by the use of the invested funds. The GDP includes final goods and services produced during the current year, and therefore, as a rule, if a good or service belongs in the GDP, then when the good or service is traded, it also belongs in the current account. For example, exports of goods and services and investment income paid to U.S. residents enter the current account as credits, while imports and foreign investment income paid by U.S. residents to foreign residents enter the current account as debits.

The capital account includes trade in assets such as international purchases and sales of bank accounts, bonds, stocks, real estate, or other international transfers of asset ownership. The purchase of U.S. assets by foreign residents, described as a *U.S. capital inflow,* is a credit on the capital account because the transaction involves a payment to U.S. residents. The purchase of foreign assets by U.S. residents, described as a *U.S. capital outflow,* is a debit on the capital account because it involves a payment made to foreign residents. Because an accommodating transaction of equal size but opposite sign must accompany every autonomous transaction, the sum of the current account balance and the capital account balance must always equal zero. For example, if the United States runs a current account deficit of $700 billion (the current account balance = –$700 billion), it must simultaneously run a capital account surplus of $700 billion (the capital account balance = +$700 billion).

## Sample Balance of Payments Transactions

Table 41.1 demonstrates the three methods for categorizing an international transaction illustrated by a U.S. resident's purchase of a Japanese automobile for $25,000. Suppose the U.S. resident completes the transaction by writing a check for $25,000, which the Japanese auto manufacturer then deposits in a U.S. bank account. The effect of these transactions on the U.S. and Japanese balance of payments is summarized here.

The originating transaction, the purchase of the Toyota automobile, is an *autonomous* transaction, accommodated by the change in foreign (Japanese) holdings of U.S. bank accounts. The capital account transaction is an

**Table 41.1** Sample Balance of Payments Transactions

| *Transaction* | *U.S. BOP* | *Japanese BOP* |
|---|---|---|
| Purchase of a Japanese automobile for $25,000 (autonomous) | Import: –$25,000 (current account) | Export: +$25,000 (current account) |
| Japanese auto manufacturer deposits the check for $25,000 (accommodating) | Capital inflow: +$25,000 | Capital outflow: –$25,000 |
| Net Effect on BOP | $0 | $0 |

*accommodating* transaction because it would not have been undertaken if not for the autonomous transaction. The import of the Japanese automobile appears on the U.S. balance of payments accounts as a debit, with a negative sign. The corresponding Japanese export appears as a credit on the Japanese balance of payments, with a positive sign. The act of depositing the $25,000 check in a U.S. commercial bank represents an increase in U.S. asset holding by foreigners, a U.S. capital inflow and a credit on the U.S. balance of payments, as well as a Japanese capital outflow and a debit on the Japanese balance of payments. The effect of the autonomous and accommodating transactions on the U.S. and Japanese balance of payments accounts is zero—in both countries, the autonomous and accommodating transactions exactly offset each other.

In practice, when the transactions shown in this illustrative example register on the balance of payments, there would be no clear connection between the import and the accommodating capital flow—the capital account transaction would appear separately from the import of the auto, and the transaction's motivation would be unknown. This makes knowing which transactions are autonomous and which are accommodating impossible, because exactly the same bank deposit as the deposit from the example could be autonomous in a different context—the foreign depositor could find the interest rate or liquidity of the deposit desirable and make the deposit independent of any other transaction on the balance of payments, which would classify the deposit as autonomous.

Many capital account transactions are autonomous; moreover, both the autonomous and accommodating transactions can occur on the capital account. Table 41.2, which illustrates a Japanese investment bank purchasing $500 million worth of stock from a U.S. corporation, illustrates an example of this possibility. Because the bank purchases the stock to earn a return, making the transaction independent of any other international transaction, the balance of payments classifies it as autonomous. The Japanese

**Table 41.2** Sample Balance of Payments Transactions

| *Transaction* | *U.S. BOP* | *Japanese BOP* |
|---|---|---|
| A Japanese investment bank purchases $500M worth of stock in a U.S. corporation (autonomous) | Capital inflow (+$500M) | Capital outflow (–$500M) |
| The Japanese investment bank pays for the stock by writing a check on its U.S. bank account (accommodating) | Capital outflow (Japanese bank sells U.S. assets by drawing $ from its account) (–$500M) | Capital inflow (Sale of U.S. assets by Japanese residents) (+$500M) |
| Net Effect | $0 | $0 |

investment bank might pay for the stock with a check, written on the Japanese bank's account in a U.S. commercial bank. This second action, drawing down the Japanese bank's deposit in the U.S. commercial bank, accommodates the purchase of the stock and would not be undertaken otherwise. On the balance of payments accounts of the United States and Japan, the autonomous and accommodating items both appear on the capital account. Again, the effect of the paired set of transactions on the accounts balance of the two countries is zero.

## Measures of the Overall Balance of Payments: The Official Settlements Balance and Others

A measure of the degree of disequilibrium in the international financial position of a country is useful to economists and policy makers, but because the comprehensive balance of payments sums to zero, due to the nature of the double-entry bookkeeping approach, any balance of payments surplus or deficit must be a subset of the comprehensive balance of payments. The *official settlements balance* (OSB) represents the net effect of changes in U.S. and foreign official reserve assets, which include foreign currency holdings, gold, special drawing rights, and the reserve position at the International Monetary Fund. Economists may use the OSB as a measure or indicator of the magnitude of balance of payments deficit and surpluses because it represents changes in the assets used by central banks for intervening in the foreign exchange market in order to peg or target the exchange rate. If a country has a balance of payments deficit, autonomous debits exceed autonomous credits, and the country's exchange authority (the central bank and possibly the country's exchange stabilization fund) must sell some of its official reserves to fill the gap. The sale of official assets appears on the capital

account as a credit. If a country has a balance of payments surplus, autonomous credits exceed autonomous debits, and the country's exchange authority must purchase foreign assets, increasing balance of payments debits, offsetting the gap between credits and debits. In a sense, the OSB provides an indicator of possible or eventual stress on the dollar because an increase in dollar-denominated asset holdings by foreign central banks can potentially be converted to foreign currency, putting downward pressure on the exchange value of the dollar. The OSB is also useful as a measure of the effect of exchange market intervention on the domestic money supply, as foreign exchange reserves are the foreign component of the monetary base, on which the money supply depends. An increase in a country's foreign exchange reserves causes an increase in the country's monetary base and its money supply.

It is important to recognize that the OSB has a limited ability to summarize a country's balance of payments position, and one must interpret the OSB with care. Central banks, like other asset holders, may find it desirable to rearrange their stock of assets, for example, to make their portfolio of assets more diverse, by purchasing or selling dollars. For example, if a U.S. resident buys a Japanese auto, and the Japanese auto firm sells the dollars it acquires to the Bank of Japan, the U.S. OSB deficit increases. Alternatively, if the Bank of Japan rearranges its assets by selling the dollars to a Japanese commercial bank, there is no change in official dollar holdings and no effect on the U.S. OSB, despite the increase in imports.

Foreign central banks frequently hold dollar-denominated reserves in institutions outside the United States, for example, in deposits at foreign commercial banks. If the foreign central bank uses these funds to intervene in the foreign exchange market, there is pressure on the exchange value of the dollar, but no effect on the OSB. Thus, the OSB may not be a dependable indicator of the extent to which central banks intervene and the degree of pressure on the dollar to appreciate or depreciate. The Bureau of Economic Analysis (BEA) *Survey of Current Business* does not explicitly state the OSB, suggesting that the U.S. government does not place a great deal of importance on the OSB as a measure of balance of payments surpluses or deficits.

During the Bretton Woods era (see Chapter 42) of fixed exchange rates and widespread central bank intervention in the foreign exchange market, two other overall balance of payments measures were common. The *basic balance* equals the sum of the current account balance and the balance on long-term capital, placing all short-term capital below the line as balancing or accommodating items. The basic balance was designed to indicate long-run tendencies in autonomous payments and receipts, but in the present day, economists rarely report or use the basic balance, mainly because they regard the division between long-term and short-term capital as arbitrary. For example, it is quite possible that international investors buy or sell short-term assets for reasons unrelated to other items on the balance of payments, so that one should classify these transactions as autonomous even though they appear on the basic balance as accommodating. The second balance of payments measure common in the Bretton Woods era, the *liquidity balance,* places short-term capital held by U.S. residents as well as the statistical discrepancy above the line, leaving short-term capital held by foreign residents as the below-the-line balancing item. While this may have some validity in the sense that short-term capital representing claims against the United States may be easy to liquidate, putting downward pressure on the dollar, the liquidity balance suffers from the same problem as the basic balance—the distinction between long-run and short-run capital is arbitrary. Foreign investors can readily liquidate some long-term capital, for example, long-term U.S. treasuries.

In the current era of managed floating, the Fed maintains relatively small foreign exchange holdings and rarely intervenes to affect the exchange value of the dollar, but a number of foreign central banks intervene heavily, so that most of the changes affecting the OSB occur on the foreign official reserve assets in the U.S. portion of the balance of payments. Foreign central banks hold the bulk of their foreign exchange reserves in the form of dollar-denominated assets, but the dollar-denominated assets held by the Fed are not viewed as foreign exchange reserves. The U.S. Treasury maintains an exchange stabilization fund that it uses for intervention along with the Fed. When the Treasury does intervene in the foreign exchange market, it uses foreign exchange reserves held in the exchange stabilization fund as well as foreign exchange reserves held by the Fed. The OSB includes changes in the foreign exchange reserves held by both institutions.

## Measures of Selected Components of the Balance of Payments, 2007

Table 41.3 shows a selection of key components of the balance of payments for the year 2007, produced by the BEA and published in the *Survey of Current Business* (www.bea.gov/international/index.htm#bop). Measures are in billions of U.S. dollars.

Exports of goods, exports of services, and investment income (income receipts) all appear in the table as credits, with a positive sign. Imports of goods, imports of services, and investment income paid to foreign residents (income payments) all appear as debits, with a negative sign. Unilateral transfers are "one-sided" transactions that include gifts from the U.S. residents to foreign residents, income earned in the United States and sent back home by foreigners working in the United States, and nonmilitary foreign aid. Although unilateral transfers are one-sided, two entries appear in the balance of payments, a debit and a credit, to preserve the double-entry bookkeeping approach. For example, if the United States donates $10 million worth of wheat as part of a foreign aid package, a $10 million debit appears on the balance

**Table 41.3** Balance of Payments, 2007

| | |
|---|---|
| Exports of goods and services and income receipts | 2,463.505 |
| Exports of goods and services | 1,645.726 |
| Goods, balance of payments basis | 1,148.481 |
| Services | 497.245 |
| Income receipts | 817.779 |
| Imports of goods and services and income payments | 3,082.014 |
| Imports of goods and services | –2,345.984 |
| Goods, balance of payments basis | –1,967.853 |
| Services | –378.130 |
| Income payments | –736.030 |
| Unilateral current transfers, net | –112.705 |
| U.S.-owned assets abroad, excluding financial derivatives (increase/financial outflow) (–) | –1,289.854 |
| U.S. official reserve assets | –0.122 |
| U.S. government assets, other than official reserve assets | –22.273 |
| U.S. private assets | –1,267.459 |
| Foreign-owned assets in the U.S., excluding financial derivatives (increase/financial inflow) (+) | 2,057.703 |
| Foreign official assets in the U.S. | 411.058 |
| Financial derivatives, net | 6.496 |
| Other capital transactions, net | –1.843 |
| Statistical discrepancy | –41.287 |
| Balance on goods | –819.373 |
| Balance on current account | –731.214 |

of payments under unilateral transfers, while a $10 million credit appears as an export of wheat, even though there has been no quid quo pro. The balance on "U.S.-owned assets abroad" refers to purchases of foreign assets, both private and official, by U.S. residents, entered as a debit on the balance of payments. The balance on "Foreign-owned assets in the U.S." represents purchases of U.S. assets by foreign residents, both private and official. The sum of these two balances equals the capital account balance.

### Measuring the Merchandise Trade Balance

The most widely used and reported subheadings of the balance of payments include the balances on the current and capital (financial) accounts, the overall trade balance, and the balance on goods, also known as the *merchandise trade balance.* The merchandise trade balance is the sum of the credit entry for goods exported and the debit entry for goods imported, which equals approximately –$819 billion for 2007. Merchandise trade is composed of exports and imports of tangible commodities, while trade in services or *invisibles* includes consulting, medical, and legal services, as well as royalties for publishing or use of intellectual property and fees for shipping costs. Separating the merchandise balance from the current account, which includes services and investment income, makes little sense conceptually, but the trade balance remains popular as a result of recording issues. While documentation of merchandise trade occurs at customs offices, so that the balance of payments can record these transactions in a fairly accurate and timely manner, transactions in services or invisibles are not so easy to discern, simply because there is no tangible good to pass through customs offices. Beyond that, services provided via the Internet or other means of sophisticated electronic communication are even more difficult to detect and record.

### Measuring the Balance on Goods and Services

The United States usually runs a surplus on the service account; for 2007, the surplus, which equals the difference between exports and imports of services, amounts to approximately $120 billion. Thus, the overall trade balance, which is the sum of the merchandise trade balance and the balance on services, equals approximately –$700 billion for 2007. The "income receipts" and "income payments" items refer to investment income, such as interest payments and dividends on assets held abroad. For example, interest payments made by domestic corporations to foreign residents appear as debits on the balance of payments, while receipts of dividends on stock owned by domestic residents appear as credits. Investment income belongs on the current account along with the balance of trade because it represents a payment for the current services provided to foreign residents by the assets. Despite the extremely large current account deficits and associated capital account surpluses (sales of domestic assets to foreign residents) for many years, investment income receipts still outweigh investment income payments for the United States.

## Current Account Balance and Capital Account Balance Measures

Perhaps the most widely used and reported component of the balance of payments is the *current account balance,* which equals the sum of the overall trade balance, the

balance on investment income, and net unilateral transfers. The current account deficit or surplus is important because it represents the change in a country's net indebtedness to foreign residents. The United States' current account balance was generally positive until the early 1980s, when the United States began steadily running current account deficits. The current account deficits have grown especially large since 2000, and for 2007, the current account stands at –$731 billion. A country must finance its current account deficit either by selling assets to foreigner residents, including borrowing, or selling assets amassed as a result of previous asset purchases from foreign residents; these transactions appear on a country's capital account as a credit, known as a *capital inflow.* Because the capital or financial account balance is identical to the current account in magnitude, but with the opposite sign, the 2007 capital account balance for the United States is +$731 billion, including the statistical discrepancy. The capital account balance equals the sum of the balances on U.S.-owned assets abroad, foreign-owned assets in the United States, net financial derivatives, and the statistical discrepancy.

While, in theory, the current and capital (financial) accounts should be of equal magnitudes, but with the opposite sign, in practice they differ, sometimes to a great degree. The statistical discrepancy bridges the difference between the actual current and capital account balances. Omitted and inaccurately recorded transactions are the primary reason for the discrepancy. The balance of payments records data collected from disparate sources at different times and, while it is impossible to know how to distribute the statistical discrepancy over the various balances included in the current and capital accounts, omitted or inaccurately recorded capital account and service transactions are the most likely sources of the discrepancy. Data collected on international asset transactions often do not capture all of the millions of daily transactions, and services are described as invisibles precisely because they are difficult to detect and document. Finally, illegal transactions and international transactions undertaken to evade taxes are, almost by definition, not documented and not included in the balance of payments. The U.S. balance of payments accounting follows the conventional method of recording the statistical discrepancy as a single line item, but there are other approaches to distributing the statistical discrepancy over the balance of payments.

## Measuring the OSB

While not given an explicit line item in the *Survey of Current Business,* the OSB equals the sum of U.S. official reserve assets (–$0.122 billion in 2007) and foreign official reserve assets in the United States ($411.946 billion in 2007); both of these balances appear in the overall capital account. As discussed in the section on classifications within the balance of payments, the OSB, while flawed, gives policy makers some idea of the size of the disequilibrium in the overall balance of payments. The OSB of approximately $411 billion indicates a gap between autonomous credits and autonomous debits of –$411 billion, suggesting that there is downward pressure on the exchange value of the dollar. As a result, foreign central banks intervene by purchasing U.S. dollars and dollar-denominated assets to prevent the dollar from depreciating and their domestic currencies from appreciating. Note that the *foreign* official reserve asset line in the U.S. balance dominates the OSB, attesting to the large-scale intervention by foreign central banks relative to the intervention by the U.S. Treasury and the Fed.

**Table 41.4** Balance of Payments, Selected Years, 1960–2005

| *Year* | *1960* | *1970* | *1980* | *1990* | *1995* | *2000* | *2005* |
|---|---|---|---|---|---|---|---|
| Exports | 30.556 | 68.387 | 344.440 | 706.975 | 1,004.631 | 1,421.515 | 1,819.016 |
| Imports | –23.670 | –59.901 | –33.377 | –759.29 | –1,080.12 | –1,780.30 | –2,458.23 |
| Unilateral transfers | –4.062 | –6.156 | –8.349 | –26.654 | –38.074 | –58.645 | –89.784 |
| U.S.-owned assets abroad | –4.099 | –8.470 | –85.815 | –81.234 | –352.264 | –560.523 | –546.631 |
| U.S. official reserve assets | 2.145 | 3.348 | –7.003 | –2.158 | –9.742 | –0.290 | 14.096 |
| Foreign-owned assets in the U.S. | 2.294 | 6.359 | 60.885 | 139.357 | 435.102 | 1,038.224 | 1,247.347 |
| Foreign official assets in the U.S. | 1.473 | 6.908 | 15.497 | 33.910 | 109.880 | 42.758 | 259.268 |
| Statistical discrepancy | –1.019 | –0.219 | 22.613 | 27.425 | 31.656 | –59.265 | 32.313 |
| Balance on current account | 2.824 | 2.331 | 2.317 | –78.968 | –113.567 | –417.426 | –728.993 |
| Official settlements balance | 3.618 | 10.256 | 8.494 | 31.752 | 100.138 | 42.468 | 273.364 |

SOURCE: Bureau of Economic Analysis, *Survey of Current Business* (http://www.bea.gov/international/index.htm#bop).

NOTE: All figures are in billions of U.S. dollars.

### Recent U.S. Balance of Payments History

Keeping in mind the caveats about the simultaneous nature of the current and capital accounts, and the difficulties with using the OSB as the overall measure of the balance of payments, there are several discernable trends that emerge from examining the U.S. balance of payments and its component balances over time.

The United States ran current account surpluses and the associated capital account deficits until the early 1980s, when the current account became increasingly negative and the capital account became increasingly positive. By 2005, the U.S. current account deficit had risen to over $700 billion, or around 5% of GDP. There are many possible explanations for this change. The price of crude oil rose in the late 1970s and early 1980s, which led to an increase in the volume of U.S. imports. In the same time frame, foreign government impediments to international capital flows were lifted, making it easier for foreign residents to satisfy their appetite for U.S. assets, leading to large U.S. capital inflows. Beyond that, U.S. national savings levels were insufficient to finance the combination of U.S. physical investment in plant and equipment and federal government deficits, necessitating increased borrowing from abroad. More recently, intervention by foreign central banks, especially China's, in the foreign exchange market to prevent their currencies from appreciating exacerbated the U.S. current account deficit by preventing a depreciation of the dollar.

The capital account surplus that accompanies the current account deficit results from the sale of assets to foreigners, including treasuries, corporate bonds, stocks, real estate, bank deposits, and so on, to foreign residents. The increased U.S. indebtedness arising from the accumulated effect of persistent current account deficits or capital account surpluses has raised alarm in some circles. The most important question regarding U.S. foreign debt is what U.S. residents use the debt to buy. If U.S. residents use the debt to purchase productive assets that yield a higher return than the interest and dividend payments to foreign residents, then U.S. society benefits from the capital account balance surplus. On the other hand, if the foreign debt finances domestic consumption, it will lead to sacrificed future consumption as U.S. residents divert future income to repay the debt.

Several features of the OSB are noteworthy. First, from 1960 until the present, the OSB was positive, indicating a net official sale of foreign assets by the United States to counteract an excess of debits over credits in the U.S. balance of payments. In the 1960s and 1970s, the negative capital account balance outweighed the positive current account balance. Starting in the early 1980s, the situation reversed, and the current account deficit outpaced the capital account surplus. With the end of the Bretton Woods system in the early 1970s, foreign official acquisition of U.S. assets dominated the OSB, mainly because the Fed and the Treasury Exchange Fund did not extensively intervene in the foreign exchange market, whereas a number of foreign central banks actively intervened.

## National Income Accounting and the Components of the Balance of Payments

### The Relationships Among the Current Account, the Capital Account, and GDP

The familiar relationship

$$Y = C + I + G + X - M, \tag{1}$$

represents the income = expenditures equilibrium condition for the national economy, where $Y$ = GDP, $C$ = personal consumption expenditure, $I$ = gross physical investment expenditure, $G$ = government purchases, $X$ = exports, and $M$ = imports. Equation 1 says that the national economy is in equilibrium if the value of production, $Y$, equals the value of expenditure for that production, $C + I + G + X - M$. Because exports are domestically produced goods and services sold to foreign residents, the expenditure side of the equation includes exports with a positive sign. On the other hand, imports are *subtracted* from the expenditure side of the expression. This occurs as a result of the way the national income and product accounts conventionally define $C$, $I$, and $G$ to *include expenditure for foreign-produced goods* as well as expenditure for domestically produced goods. For example, personal consumption, $C$, includes not only expenditure by domestic consumers for domestically manufactured automobiles, but also expenditure by domestic consumers for automobiles manufactured in Japan, Korea, and Germany, for example. Investment expenditure includes purchases of new plant and equipment produced domestically, as well as purchases of new plant and equipment produced abroad. The income and product accounts define government purchases similarly.

Subtracting domestic expenditure, $C + I + G$, from both sides of the equilibrium condition yields

$$Y - (C + I + G) = X - M. \tag{2}$$

$X - M$, broadly defined to include investment income as a payment for a current service and unilateral transfers, equals the current account balance. Thus, rewriting the equation this way suggests that the difference between national income and national expenditure equals the current account balance. If income exceeds expenditure by a given amount, then exports exceed imports by the same amount and there is a current account surplus:

$$Y - (C + I + G) > 0 \text{ implies } X - M > 0.$$

If national expenditure exceeds national income, then imports exceed exports by the same amount and there is a current account deficit:

$$Y - (C + I + G) < 0 \text{ implies } X - M < 0.$$

For example, if U.S. output generates $14,000 billion in national income, and total expenditure by domestic residents

for currently produced goods and services is $14,700 billion, then there is an additional expenditure of $700 billion for goods and services produced abroad because spending for domestic output must be limited to $14,000 billion. That is, the United States buys $700 billion more in goods and services from foreign residents than foreign residents buy from the United States, and the United States runs a current account balance of –$700 billion.

As a result of the double-entry bookkeeping approach, any transaction, whether classified on the balance of payments as a credit or debit, must have an associated transaction of equal size and opposite sign, suggesting that the sum of the current account and the capital account must always be zero. As discussed in the previous sections on the components of the balance of payments, if the United States has a current account deficit of $700 billion ($X - M = -\$700$ billion), then it must have a capital account surplus of $700 billion ($M - X = +\$700$ billion).

Multiplying both sides of the equation $Y - (C + I + G) = X - M$ by –1 yields

$$(C + I + G) - Y = M - X. \tag{3}$$

In the form of Equation 3, the equilibrium condition suggests that the difference between domestic national expenditure and domestic national income equals the capital account balance:

$$(C + I + G) - Y = \text{capital account balance}.$$

Thinking about how it is possible for domestic expenditure to be greater than domestic income bolsters this idea. Like an individual, the only way a country can spend more than the income it generates is for the country to sell assets; these assets could include bank accounts, real estate, corporate stocks, certificates of indebtedness, and so on. Thus, in order for a country to spend more than its national income, it must sell more assets to the rest of the world than it buys from the rest of the world; in other words, the country must run a capital account surplus. A sale of assets results in a receipt from foreign residents and, as a result, appears as a credit on the capital account. Conversely, we can say that the only way an individual, business firm, or nation can sell more assets to the rest of the world than it buys from the rest of the world is to buy more goods from the rest of the world than it sells to the rest of the world.

## The Current Account and the Domestic Savings–Investment Relationship

There are only three things households can do with the income generated by the production of GDP; after they have paid taxes "off the top" of their income, households can either spend or save the remainder, usually defined as *disposable income,* implying that the expression $Y = C + S + T$, where $S$ = private savings and $T$ = domestic tax revenue for all levels of government, must always hold. Because $Y$ equals both $C + S + T$ and $C + I + G + X - M$, it follows that

$$C + S + T = C + I + G + X - M. \tag{4}$$

Subtracting $C$ from both sides of Equation 4 yields another expression of national economic equilibrium,

$$S + T = I + G + X - M. \tag{5}$$

Rearranging terms in Equation 5 by adding $(M - X)$ to both sides and subtracting $T$ from both sides provides one useful approach for discussing the connection between the current account and the savings–investment relationship.

$$S + (M - X) = I + (G - T) \tag{6}$$

Equation 6 says private savings plus the capital account balance, shown on the left-hand side, must finance investment and the government budget deficit, shown on the right-hand side.

A country's capital account balance is, in essence, the country's net sale of assets to foreign residents. A capital account surplus of $700 billion denotes that U.S. residents sell assets to foreign residents worth $700 billion more than U.S. residents buy from foreign residents and, as a result, there is a net capital inflow that supplements private savings to fund U.S. investment and the U.S. government budget deficit. Approximating the data provided by the National Income and Produce Accounts for 2007, U.S. private savings is $1900 billion, and the U.S. capital account balance is +$700 billion, which allows the United States to finance $2400 billion in investment and $200 billion in government borrowing as shown in Figure 41.1.

The relationship shown here between the capital account balance and the government budget deficit gives rise to the discussion of the *twin deficit* phenomenon.

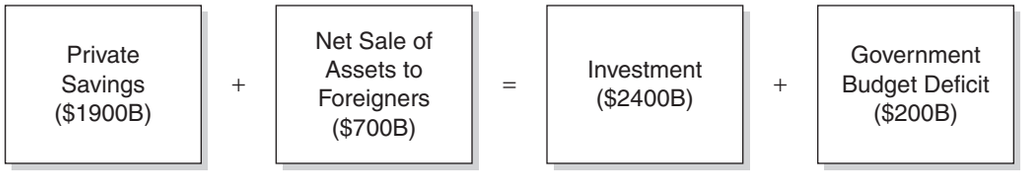


**Figure 41.1** Capital Account and the Savings–Investment Relationship

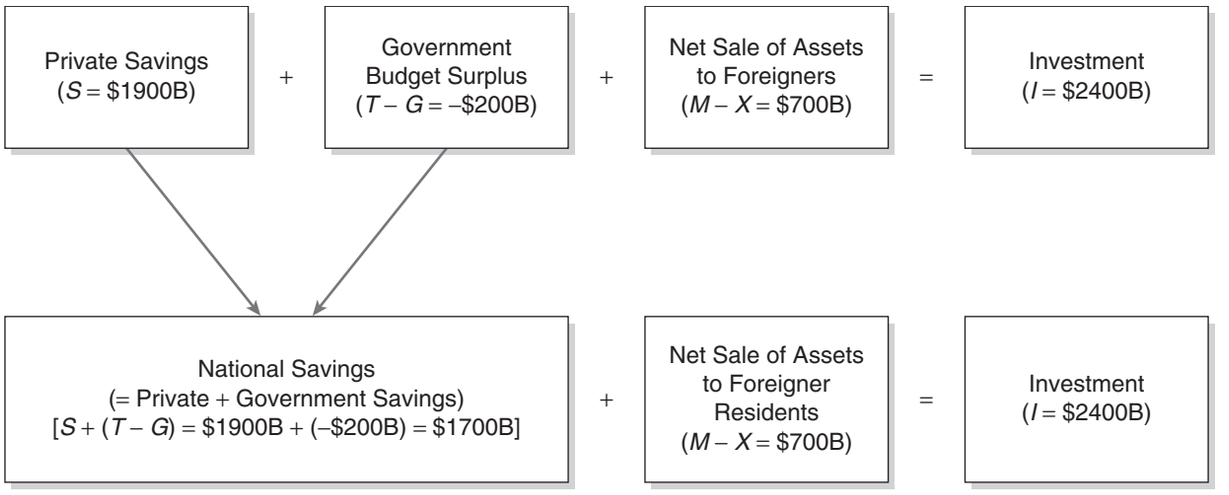


**Figure 41.2** National Savings and the Capital Account

With private savings and domestic investment held constant, an increase in the budget deficit causes an increase in the net sale of assets to foreign residents, so that the capital account surplus, hence the current account deficit, increases.

Alternatively, it is possible to rearrange the terms in Figure 41.1 to see the relationship between *national savings* and the current or capital account balance by moving the government budget relation to the left-hand side of the equation to obtain Figure 41.2.

$S + (T - G)$ represents national savings, the sum of private savings and the government budget surplus, which represents government savings. Figure 41.2 says the sum of national savings, $S + (T - G)$, and net capital inflows, $(M - X)$, provides the funding of private national investment—by selling assets to foreign residents, the economy can finance more investment than it could if the economy were closed, ceteris paribus.

## Balance of Payments and the Foreign Exchange Market

Figure 41.3 illustrates the foreign exchange market, exemplified by the market for British pounds priced in terms of dollars. The demand for foreign currency, shown in Figure 41.3, derives from domestic residents' autonomous

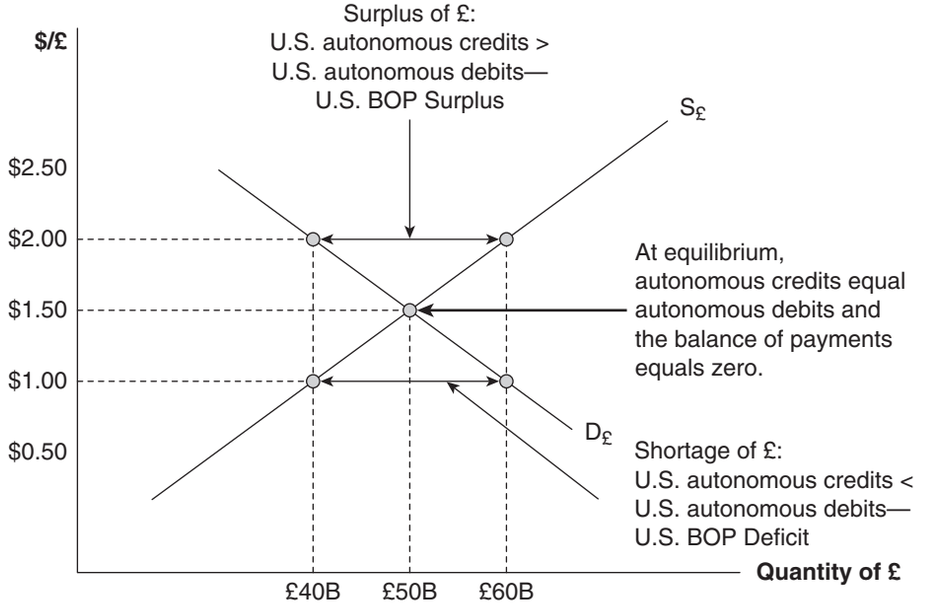


**Figure 41.3** The Foreign Exchange Market

purchases of foreign goods, services, and assets, and the supply of foreign currency derives from foreign residents' autonomous purchases of domestic products and assets. When the foreign exchange market is in equilibrium, shown in Figure 41.3 at the exchange rate of $1.50 = £1, autonomous credits equal autonomous debits and the balance of payments equals zero. When the foreign exchange market is out of equilibrium, autonomous credits and debits are not equal and there is either a balance of payments surplus or deficit.

### Floating Exchange Rates

Under a floating exchange system, the exchange rate moves freely to equalize autonomous credits and autonomous debits so that the autonomous quantity demanded and the autonomous quantity supplied of pounds coincide at the equilibrium exchange rate of $1.50 = £1. As a result, there is neither an excess demand nor an excess supply of foreign currency, indicating that the balance of payments equals zero under floating exchange rates.

### Fixed Exchange Rates

If a country fixes or "pegs" the exchange rate at a target level, the exchange rate cannot move to equate the demand and supply of the currencies traded on the foreign exchange market, resulting in an excess demand or an excess supply of the foreign currency, which the country's central bank must satisfy to maintain the peg. Figure 41.3 illustrates the effect on the balance of payments of the two general cases, one in which the central banks pegs the exchange rate above the equilibrium of $1.50/£, and the other in which the central banks pegs the exchange rate below the equilibrium rate.

Where the central bank pegs the exchange rate at $2.00/£, there is an excess supply of pounds offered on the foreign exchange market, representing a U.S. balance of payments surplus of £20 billion. The balance of payments surplus indicates that at $2.00/£, U.S. residents want to spend less on British products and assets than British residents want to spend for U.S. products and assets. In order to maintain the exchange rate at its targeted level, and prevent the pound from depreciating, the Federal Reserve and the Bank of England must purchase the excess supply of pounds with dollars. The extent of the intervention, measured by the cost of the pounds purchased, equals the U.S. balance of payments surplus. If the intervention amounts to $40 billion, for example, the U.S. balance of payments exhibits a debit of $40 billion that compensates for the excess of credits over debits on the U.S. balance of payments created by the autonomous transactions.

Where the central bank pegs the exchange rate below equilibrium, at $1.00/£, there is an excess demand for pounds, representing a U.S. balance of payments deficit; at $1.00/£, U.S. residents demand more pounds than British residents offer, indicating that at $1.00/£, U.S. residents want to spend more on British products and assets than British residents want to spend for U.S. products and assets. In this case, the Federal Reserve and the Bank of England must satisfy the excess demand for pounds that exists at $1.00/£ by selling pounds on the foreign exchange market in exchange for dollars. The sale of pounds or pound-denominated assets reduces the Fed's holdings of foreign exchange reserves and registers on the U.S. balance of payments as a credit equal to the extent of the intervention. Again, the intervention compensates for the excess of autonomous debits over autonomous credits on the U.S. balance of payments.

## Conclusion

The balance of payments documents a country's international financial position, recording international transactions in goods and services as well as international transactions in assets. The balance of payments not only takes account of private transactions, but also encompasses official transactions that central banks make when they intervene in the foreign market. Balance of payments accounting relies on a double-entry bookkeeping approach in which every transaction appears twice, once as a credit and once as a debit. The balance of payments also classifies every international transaction as either autonomous or accommodating. Traders make autonomous transactions independent of the balance of payments, initiating accommodating transactions that would not otherwise occur. Conceptually, a country's *payments balance* is the sum of all autonomous credits and debits. This sum could be positive, representing a balance of payments surplus, or negative, representing a balance of payments deficit. In practice, however, it is not possible to distinguish autonomous and accommodating transactions, so that no measure of the balance of payments is without its shortcomings.

As a result of the double-entry bookkeeping approach, the overall balance of payments, which includes all credits and debits, whether they are autonomous or accommodating, always sums to zero. Thus, rather than focusing on the overall balance of payments, policy makers usually focus on its component balances. The most important of these is the current account balance, which includes exports and imports of goods and services, as well as investment income and unilateral transfers. The current account balance measures the change in a country's international indebtedness. If a country runs a current account deficit, it must sell assets to the rest of the world, running a capital account surplus of equal size. Selling assets to the rest of the world is not necessarily harmful for a country. If the return to the country on the foreign assets outweighs the interest, dividends, and so on, that the county's residents

have to pay to foreign residents, the sale of assets improves the county's welfare. There is also a strong link between a country's domestic savings–investment relationship and the country's current account balance. If a country's domestic savings is not sufficient to fund its domestic physical investment, the country must borrow from abroad, resulting in a capital account surplus and a current account deficit.

The foreign exchange market relates closely to the balance of payments. In the foreign exchange market, the demand and supply of currency represent only autonomous transactions, so that when the market is in equilibrium under a freely floating exchange rate, autonomous credits and debits are equal and the balance of payments is also in equilibrium. If the country targets or pegs the exchange rate, the resulting excess demand or excess supply of currency represents a balance of payments deficit or surplus.

## References and Further Readings


Destler, I. M., & Randall Henning, C. (1990). *Dollar politics: Exchange rate policymaking in the United States.* Washington, DC: Institute for International Economics.

Faust, J. (1989). U.S. foreign indebtedness: Are we investing what we borrow? *Federal Reserve Bank of Kansas City Economic Review, 74*(7), 3–20.

Friedman, M. (1953). The case for flexible exchange rates. In *Essays in positive economics* (pp. 157–203). Chicago: University of Chicago Press.

Graboyes, R. F. (1991, September–October). International trade and payments data: An introduction. *Federal Reserve Bank of Richmond Review, 77,* 20–31.

Higgins, M., & Klitgaard, T. (1998, December). Viewing the current account deficit as a capital inflow. *Federal Reserve Bank of New York Current Issues in Economics and Finance, 4,* 1–6.

Hooper, P., & Richardson, J. D. (1991). *International economic transactions: Issues in measurement and empirical research.* Chicago: University of Chicago Press.

International Monetary Fund. (Yearly). *Balance of payments statistics yearbook.* Washington, DC: Author.

Kemp, D. S. (1975, July). Balance of payments of concepts—What do they really mean? *Federal Reserve Bank of St. Louis Review, 57,* 14–23.

Meade, J. E. (1951). *The theory of international economic policy: Vol. 1. The balance of payments.* London: Oxford University Press.

Schumpeter, J. A. (1954). *History of economic analysis.* New York: Oxford University Press.

Stekler, L. (1990). Adequacy of international transactions and position data for policy coordination. In W. H. Branson, J. A. Frenkel, & M. Goldstein (Eds.), *International policy coordination and exchange rate fluctuations* (pp. 347–371). Chicago: University of Chicago Press.

Stern, R. M. (1973). *The balance of payments: Theory and economic policy.* Chicago: Aldine.

U.S. Department of Commerce, Bureau of Economic Analysis. (1990, May). *The balance of payments of the United States: Concepts, data sources, and estimating procedures.* Washington, DC: Government Printing Office.

What drives large current account deficits? (2001, May). *Federal Reserve Bank of St. Louis International Economic Trends,* p. 1.

# 42

# Exchange Rates

Carlos Vargas-Silva
*Sam Houston State University*

Every day, millions of companies and individuals around the world do business with companies and individuals located in different countries. Often these transactions involve two or more currencies. Juanita, a resident of the United States, wants to send her grandmother in Mexico money to buy a new microwave. Juanita earns income in U.S. dollars, but her grandmother needs Mexican pesos to buy the microwave in Mexico. Juanita knows that the microwave costs 20,000 Mexican pesos. How many U.S. dollars would she need to send to her grandmother? This example is just one of many situations in which there is a need for a foreign exchange transaction. Individuals and companies enter into millions of foreign exchange transactions every day. The standard for how many units of one currency (e.g., U.S. dollars) must be surrendered to obtain one unit of another currency (e.g., Mexican pesos) is commonly known as the *exchange rate* between two currencies. That is, the exchange rate is the price of one currency in terms of another currency.

The value of the currency has huge implications for a nation's economy. If the value of the domestic currency appreciates—that is, more units of the foreign currency are needed to buy one unit of the domestic currency—then this country will be able to buy foreign products more cheaply. On the other hand, the products of this country are going to be more expensive in foreign markets, and as a result, the stronger currency may hurt local exporters as foreign demand decreases. It can be argued that not only the value of a currency but also its predictability can affect international businesses. If one company is going to have operations in a country whose currency has an uncertain future value, then there may be uncertainty about this company's future revenues, operational costs, and therefore profits. As it is well known, risk-averse companies prefer to decrease the risk and uncertainty of future revenues and costs.

In addition to its importance for trade and investment across countries, the exchange rate is important as the essential component of the foreign exchange market or "FX" market. *The FX market,* the global market for currencies, is the largest and most liquid financial market in the world. According to a survey conducted by the Bank for International Settlements (2002), the average daily trade in the FX market surpasses U.S.$1.2 trillion.

All of these reasons and many more make study of exchange rates an essential component of the curriculum for any student of economics or aspiring economist. That is why this handbook has decided to dedicate an entire chapter to the topic of exchange rates. In this chapter, you will learn about the functioning of the foreign exchange rate system. We start by giving a historical account of the different exchange rate systems around the world. In the past, particularly in the period from World War II to the end of the Bretton Woods era, it was common for many countries around the world to have fixed exchange rates. That is, the value of the currency was pegged to the value of some other currency. This arrangement was adopted by different countries with the hope that it would avoid uncertainty about the value of their currency and the economic consequences of such uncertainty. Nowadays, most major currencies are flexible, and their values are determined by demand and supply in the foreign exchange market.

Next, we discuss the determinants of a currency's value according to the main theories of exchange rate determination. The determinants of the value of a currency are far from settled in the academic literature. However, several theories of exchange rate determination provide useful

insights about the functioning of the foreign exchange rate market. For example, by looking at theories of exchange rate determination, we can relate the exchange rate to other important macroeconomic variables such as interest rates and prices. Once we have a good understanding of the history and determinants of the exchange rate, we discuss the difference between the nominal and real exchange rate. Finally, the chapter ends with conclusions and a list of additional readings on the topic.

## Exchange Rate Regimes: A Historical Perspective

Let us start the discussion by defining three main types of exchange rate regimes differentiated by their degree of flexibility. First, there is the floating exchange rate regime in which currencies are free to float and their values are determined by demand and supply in the foreign exchange rate market. Proponents of the flexible exchange rate system argue that floating exchange rates provide monetary policy autonomy as the central bank is not required to intervene in the foreign exchange market to maintain a given value for the currency and insulation from external shocks. In his 1969 article praising the benefits of floating exchange rates, economics professor Harry G. Johnson even argued that floating exchange rates are an essential component of the national autonomy and independence of each country. This national autonomy, he argues, is essential for the efficient organization and development of the global economy.

On the down side, flexible exchange rates have often been associated with destabilizing speculation. However, proponents of flexible exchange rates argue that, to the contrary, speculation by rational agents should reduce exchange rate volatility. Economist and Nobel laureate Milton Friedman (1953) argues that if there is a domestic currency appreciation that is expected to be temporary, there is an incentive for holders of domestic currency to sell some of their holdings, acquire foreign currency, and then buy the domestic currency back at a lower price. By behaving this way, speculators meet part of the excess demand for domestic currency and accelerate the process of returning to the long-term equilibrium value of the currency. Friedman stated that those who argue that speculation can be destabilizing in the foreign exchange market do not realize that this is equivalent to saying that speculators constantly lose money because speculation can be destabilizing only if speculators sell when the currency is cheap and buy when it is expensive. He suggests that speculators who behave this way will be driven out of the market relatively quickly.

The second type of regime that must be defined is the fixed exchange rate regime, in which the value of the currency is tied to the value of some other currency. The exchange rate uncertainty that may result from floating exchange rates has a negative impact on some types of investments, possibly affecting trade. It is believed that by fixing the value of the domestic currency relative to that of a major economy, a country can lower exchange rate volatility and promote trade and investment. Fixed exchange rates have been seen by some economists as helpful in obtaining several other economic goals. For instance, it has been argued that fixed exchange rates are useful in obtaining price stability. An exchange rate target is straightforward and easily understood by the general public. If this exchange rate target is credible—that is, if the public has confidence that the monetary authorities will pursue such a target—then it may lower inflation expectations to the level prevailing in the anchor country.

Some fixed exchange rate arrangements imply the surrender of the monetary authorities' control over domestic monetary policy. For some countries (especially developing countries), with a lack of discipline and too many incentives to create revenue from money creation, a surrender of the monetary policy may be a desirable outcome. Even Milton Friedman, who championed floating exchange rates, admitted that fixed exchange rates can be sometimes preferable for developing countries. However, because fixed exchange rate arrangements limit the monetary policy options of the country, these arrangements can expose the country to international shocks.

Some examples of fixed exchange rate regimes include dollarization and currency boards. In dollarized economies, the currency of another country circulates as the sole legal tender. The term *dollarization* refers to any country that uses a foreign currency as the domestic currency, not only to those countries that adopt the U.S. dollar. Dollarization differs from a monetary union in which members belong to a currency union in which the same legal tender is shared by members of the union. The adoption of a dollarized regime implies the complete surrender of the monetary authorities' control over domestic monetary policy. In the case of currency boards, there is an explicit legal commitment to exchange domestic currency for a specified foreign currency at a certain value, combined with restrictions on the monetary authorities' power to ensure the execution of that legal obligation.

Finally, our third type of regime is the intermediate regime. This regime is essentially some type of pegged float in which the exchange rate is free to fluctuate but is kept by the country's monetary authorities from deviating from a certain range. There are numerous intermediate arrangements, and the possibilities are limited only by the imagination. Some of the intermediate exchange rate regimes recognized by the International Monetary Fund (IMF) include pegged exchange rates within horizontal bands, crawling pegs, exchange rates within crawling bands, and managed floating with no predetermined path for the exchange rate. Next we discuss the International Monetary Fund definition of each of these regimes.

The pegged exchange rate within horizontal bands is a regime in which the value of the currency is maintained within a range of 1%. In a crawling peg regime, the exchange rate is adjusted periodically at a fixed rate or in response to changes in several indicators (e.g., relative inflation measures). These two regimes leave the country with a limited degree of monetary policy discretion. The exchange rate within crawling bands is a regime that

combines certain aspects of the previous two regimes. Finally, in a regime of managed floating with no predetermined path, the monetary authority influences the exchange rate without having a specific exchange rate path or target.

Across time, countries have moved from one currency regime to another to facilitate business and improve national economies. Next we provide a short historical recount of the main trends in exchange rate regimes.

## The Gold Standard

Although there were some reappearances of a gold standard after 1914 and the outbreak of World War I, it is agreed that 1914 marks the end of the classical gold standard era, during which the majority of countries adhered to some type of gold standard system. Under the gold standard, currencies were linked to gold—that is, the value of a country's currency was set at a given rate to gold ounces.

Gold has been traditionally used for commercial purposes as a medium of exchange and store of value because it is rare and durable. Paper currency itself has no intrinsic value (it is just plain paper) but was accepted as payments for goods and services because it could be redeemed any time for its equivalent value in gold. The gold standard also worked as an international pegged exchange rate system. Countries maintained a fixed price for gold, and therefore, if we take into account transaction costs, the rates of exchange between currencies were limited by a very tight band. The automatic adjustment process under the gold standard implied that sometimes policy tools were not available to be used for fighting cyclical economic downturns or inflation.

The gold standard regulated the quantity and growth rate of a country's money supply. This, according to economics professor and gold standard expert Michael D. Bordo (1981, 1993), implied that the price level should not vary much in the long run. That is, under the gold standard, there would be a tendency toward price stability. In 1914, with the advent of World War I and the financing of war-related expenses by printing money, the classical gold standard era came to an end.

## The Bretton Woods Period

After 1914, there were some efforts to establish a new gold standard that for several reasons, including the Great Depression, were unsuccessful. Many countries decided to abandon the gold standard and adopt free-floating exchange rates.

At the close of World War II, delegates from the Allied nations gathered at the Mount Washington Hotel in Bretton Woods, New Hampshire, in the United States. Delegates deliberated extensively in a effort to establish the basic rules of the international monetary system. The goals of the attendees included global economic stability and increased volumes of global trade. The agreements included the creation of the International Bank for Reconstruction and Development (IBRD; now part of the World Bank Group) and the IMF. Attendees also agreed that there were disadvantages in the use of floating exchange rates. It was, therefore, also agreed that currencies would once again be fixed. World major currencies were pegged to the U.S. dollar. At this point, the U.S. dollar was pegged to gold at $35 per ounce of gold. The Bretton Woods system lasted until the early 1970s, when the U.S. dollar could no longer hold the peg of $35 per ounce of gold and the convertibility of U.S. dollars to gold was suspended.

## The Post–Bretton Woods Period

After the Bretton Woods era ended, countries moved in several directions. Several developed countries, such as the United States, have since embraced floating exchange rates. These countries determined that exchange rates were no longer the optimal method to conduct monetary policy. It has become the convention to argue that exchange rates should be determined directly by market forces and should be free to fluctuate continually. Many developing countries have followed suit and have liberalized their currencies.

However, it is argued that many countries that say they allow their exchange rate to float are not really floaters and intervene frequently in the foreign exchange market to manipulate the value of their currency. Economists Guillermo A. Calvo and Carmen M. Reinhart (2000) argue that because countries that are classified as having a free float mostly resemble noncredible pegs, the demise of fixed exchange rates is a myth. They argue that many of these supposed floaters suffer from an epidemic of "fear of floating."

On the other hand, another series of countries have explicitly decided to adopt fixed or pegged exchange rate systems. Many of these countries have arrangements that fall in one of two categories. The first category includes countries that favor regional arrangements. For instance, during the 1970s, there was a regional movement toward currency integration in Europe that resulted in the European Monetary System (EMS) in 1979, a pegged exchange rate regime within horizontal bands. According to Bordo (1981, 1993), the motivation for the EMS included the strong dislike by Europeans of flexible exchange rates and the common agricultural policy established in Europe in 1959. Later on it was decided to replace most of the region's currencies for a single currency. The final result was the euro (€), the official currency of the European Union (EU). The euro was launched on January 4, 1999, with 11 member states of the EU and is currently used by 15 member states (Austria, Belgium, Cyprus, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, the Netherlands, Portugal, Slovenia, and Spain).

Other countries have decided to dollarize their economies. These countries have decided to disregard discretionary monetary policy and use a foreign currency instead of the domestic currency. Some examples of recently dollarized economies include Ecuador and El Salvador. In both cases, the U.S. dollar is the sole legal tender.

Given all these examples, someone may ask which currency regime is the best. Should countries adopt flexible exchange rates? Are countries better off with a fixed currency? Are intermediate regimes the best alternative?

Bordo (1981, 1993) and Frankel (1999), among other economists, argue that no single currency regime is optimal for all countries. The best regime depends on the specific case of each country, and a case can be made for each regime under particular circumstances.

## Determinants of the Exchange Rate

Economists have not yet reached a consensus on the factors that should be included as determinants of the exchange rate. Exchange rates may respond to many different variables, some of which are real variables (e.g., exports and imports) and some of which are financial and monetary variables (e.g., interest rates and inflation). Let us start the discussion of the determinants of the exchange rate with a simple model of exchange rate determination in which the exchange rate is determined by the interaction of supply and demand for foreign exchange. We assume that domestic and foreign exporters are paid in their respective country's currency. For example, Japanese exporters are paid in yen, while American exporters are paid in U.S. dollars. Both the supply and demand for foreign exchange are determined by the amount exported and imported in each country. For instance, if Americans want to import Japanese cars, then they must pay the Japanese exporters for their cars using yen. Alternatively, if Japanese tourists want to visit New York, they need U.S. dollars for expenditures in food, accommodations, entertainment, and shopping. These expenditures are considered U.S. exports and are a source of demand for U.S. dollars.

The interaction between demand and supply in this market is represented in Figure 42.1. We label domestic exports as $X$, domestic imports as $M$, domestic prices as $P$, and foreign prices as $P^*$. The demand for foreign currency is the foreign price multiplied by the amount imported, that is, $(P^*)(M)$. In our previous example, this will be the number of Japanese cars multiplied by the price of those cars. The supply of foreign exchange is the domestic price (adjusted by the exchange rate) multiplied by the amount exported, that is, $(P/S)X$. If a country increases exports, then there is an increase in the supply of foreign exchange as foreigners buy domestic currency to pay for the exports. On the other hand, imports increase the demand for foreign exchange as more foreign currency is needed to pay for the foreign goods. The equilibrium between demand and supply is point $E$, and the exchange rate that equilibrates demand and supply is $S_E$.

Suppose that there is an increase in the demand for foreign exchange. As presented at the bottom of Figure 42.1, as a result of the increased demand for foreign exchange, there is a shift of the demand curve from $DE$ to $DE'$ and a depreciation of the domestic currency. The new equilibrium is point $b$. At this point, more units of the domestic currency are needed to buy one unit of the foreign currency. If the monetary authority of the domestic country wants to avoid the domestic currency depreciation, then it must shift the supply curve for foreign exchange from $SU$ to $SU'$, getting to the new equilibrium point $c$ at which the exchange rate gets back to $S_E$.

### The Forward Exchange Rate, Swaps, Futures, and Options

Until now we have been analyzing the current rate of exchange between two currencies, also known as the *spot exchange rate*. That is, we have been referring to the rate of a foreign exchange contract that is traded today for immediate delivery. In contrast, there is the rate of a foreign exchange contract that is traded today but for delivery on a

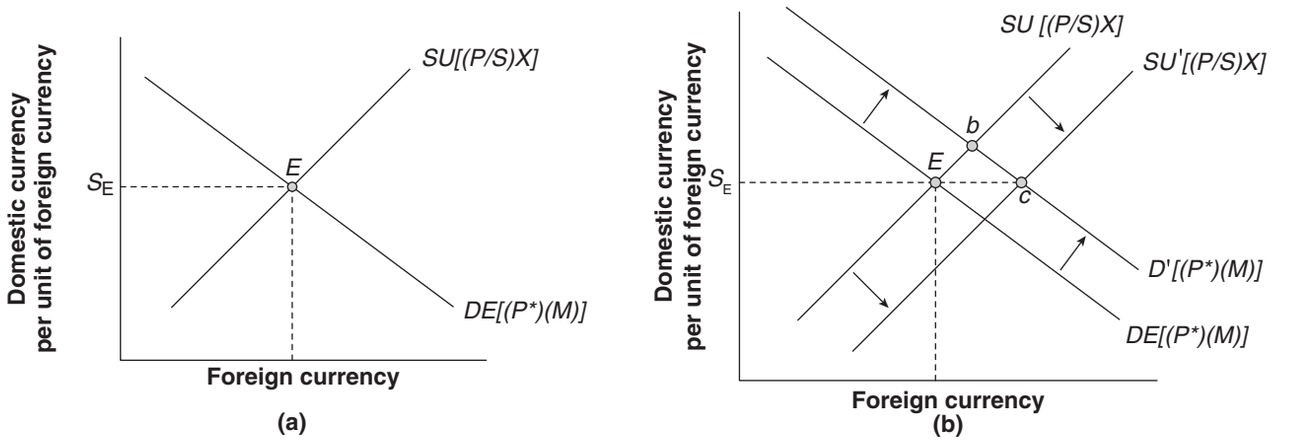


**Figure 42.1** A Simple Model of Exchange Rate Determination: (a) Foreign Exchange Rate Market Equilibrium; (b) Increase in Demand for Foreign Exchange

future date, which is what economists refer to as the *forward exchange rate.* The forward exchange rate is a hedging tool against exchange rate risk. By locking the rate at which the currency is going to be exchanged in the future, companies and individuals avoid the risk associated with making payments in other currencies at a future date.

Let us use an example to explain why individuals and businesses around the world may wish to engage in forward exchange transactions. For instance, imagine that a domestic car importer knows that he must pay yen to a Japanese car exporter for 10 Toyotas in 6 months. The domestic car importer sells each car for $10,000 and must pay the Japanese exporter ¥900,000 per car. Therefore, the domestic car importer's profit from selling the cars depends on the U.S. dollar/Japanese yen exchange rate. The current U.S. dollar/Japanese yen exchange rate is U.S.$0.0094 per yen, and the 180-day forward U.S. dollar/Japanese yen exchange rate is U.S.$0.0098 per yen. Therefore, if the domestic car importer were to pay the Japanese exporter today, he would have to pay per car a total of U.S.$8,460 (¥900,000 × $0.0094 = $8,460). But remember that the transaction is going to take place in 6 months. Imagine what will happen if, in 6 months, the U.S. dollar/Japanese yen spot rate is U.S.$0.012 per yen. In this case, the domestic importer has to pay the Japanese exporter U.S.$10,800 per car (¥900,000 × $0.012 = $10,800). He is going to sell each car for U.S.$10,000 and therefore is losing U.S.$800 on each car. One way for the domestic importer to avoid this risk is to use the forward market. He can make a contract to buy yen in 180 days at a rate of $0.0098, and therefore he will be paying the Japanese exporter $8,820 per car, making a profit of $1,180 per car.

It is also possible for companies to combine a spot sale of currency with a forward purchase of the currency in just one transaction, potentially lowering their costs. For instance, suppose that the car domestic importer has ¥900,000 today. Knowing that he has to pay yen in 6 months to the Japanese exporters, the domestic importer agrees to convert the ¥900,000 into U.S. dollars today and reconvert those into yen in 6 months. In the meantime, the domestic importer can invest the U.S. dollars in interest-generating assets (e.g., bonds) in the United States. This type of transaction, in which there is a spot foreign exchange transaction and a forward foreign exchange transaction, is called a *foreign exchange swap.* The spot and forward transactions are called the *legs of the swap.*

Two other financial instruments related to the foreign exchange market that are commonly used are futures contracts and foreign exchange options. A *futures contract* is a promise that a specific amount of foreign currency will be delivered on a specific future day. Different from a forward deal, investors can sell futures contracts in an organized market. A *foreign exchange option* is an option to buy or sell a specified amount of foreign currency at some specified date at a specified price. If the owner of the option exercises his or her right to buy the currency, then we say that he or she calls the option. The owner of the option has the right, but not the obligation, to exchange one currency for another currency at the specified date and price.

## Covered Interest Rate Parity

The forward exchange rate may also be used by portfolio managers who want to invest in other countries and earn a return that is free of exchange rate risk. The *covered interest parity condition* (CIP) requires an investor to be indifferent between placing an extra dollar in domestic or foreign investments if the rate of returns is equal and risk free. This concept can be quite easily represented with the following equation:

$$1 + i = \frac{(1 + i^*)F}{S}, \tag{1}$$

where $i$ ($i^*$) refers to the domestic (foreign) rate of return on investments, $F$ is the forward exchange rate, and $S$ is the spot exchange rate. An investor is indifferent between investing in domestic bonds and investing in bonds in the foreign country if there is certainty about the future value of the foreign investment in terms of domestic currency and the returns on both investments are equal. For the future return to be certain in domestic currency, the investor buys foreign currency at price $S$ in order to buy bonds in the foreign country and then "covers" the investment by means of the forward exchange rate $F$. If the returns to the investments are different (the left-hand side of Equation 1 is different from the right-hand side), then investors could take advantage of this difference in return and make risk-free profits. It would be possible for an investor to borrow money in the country with the lower interest rate and invest the money in the country with the higher interest rate. The possibility of taking advantage of price or return differences on different markets is commonly known as arbitrage. Any such possibility of arbitrage will be corrected quickly by the market, returning us to equality in Equation 1.

We can rearrange Equation 1 to obtain

$$i - i^* = \frac{F - S}{S}$$

or also

$$i - i^* = p, \tag{2}$$

where $(F - S)/S = p$. The term $p$ is known as the forward premium of foreign currency against domestic currency

and is the cost of covering the transaction. The domestic interest rate equals the interest rate on foreign deposits plus the forward premium. When the domestic interest rate is below the foreign interest rate ( $i < i^*$ ), the forward price of the foreign currency is below the spot price. Conversely, if the domestic interest rate is above the foreign interest rate ( $i > i^*$ ), the forward price of the foreign currency exceeds the spot price.

Economists Jacob A. Frenkel and Richard M. Levich (1975) tested CIP and found that in many instances, there was a covered interest rate differential. Is this covered interest rate differential evidence of unexploited profit opportunities? Not really, according to Frenkel and Levich. They argue that after taking into account transaction costs, most of the empirical deviations from CIP disappear.

## Uncovered Interest Rate Parity

There is another form of interest rate parity known as uncovered interest rate parity (UIP). In this case, a risk-neutral investor is indifferent between placing an extra dollar in domestic or foreign investments if the expected rates of return are equal. In this case, the foreign investment is not risk free. This concept can be represented with the following equation:

$$1 + i_t = \frac{(1 + i_t^*)ES_{t+n}}{S_t}. \tag{3}$$

There are two main differences between Equation 1 and Equation 3. First, we are adding time subscripts to the variables. This is because now we will be referring to variables in different time periods. Second, we are substituting $ES_{t+n}$ for $F$, where $ES_{t+n}$ refers to the expected spot exchange rate in $n$ periods. Under this condition, an investor is indifferent between investing in domestic bonds and investing in foreign bonds with a similar expected return. The investor buys foreign currency at price $S$ to buy bonds in the foreign country and then return those investments to domestic currency in $n$ periods at a rate that he or she thinks will be equal to $ES_{t+n}$. However, there is no certainty about the future value of the spot rate, and hence the investor is subject to exchange rate risk. We say that his or her transaction is "uncovered."

We can use a common approximation to UIP that assumes, among other things, that

$$\frac{ES_{t+n}}{S_t} = \Delta ES_{t+n},$$

where $\Delta ES_{t+n}$ represents the expected change in the exchange rate, to obtain the interest rate differential,

$$i_t - i_t^* = \Delta ES_{t+n}. \tag{4}$$

Equation 4 relates the interest rate differential to the exchange rate. When the domestic interest rate is below the foreign interest rate ($i_t < i_t^*$), the domestic currency is expected to appreciate in the future. Hence, individuals who invest money in the domestic market today at a lower interest rate expect to be compensated in the future by the appreciation of the domestic currency. Similarly, if the domestic interest rate is above the foreign interest rate ( $i_t > i_t^*$ ), the domestic currency should be expected to depreciate in the future. The only reason someone will invest in the foreign country (assuming no transaction costs and a similar level of risk) with a lower interest rate is if he or she expects a depreciation of the domestic currency with respect to the foreign currency.

It is not quite as easy to test empirically for UIP as it is to test for CIP, mainly because it is not possible to observe expected exchange rates. Therefore, the researcher must make an assumption about the formation of agents' expectations about the future value of the currency. Often researchers testing the empirical validity of UIP assumed that expectations are formed rationally. Rational expectations assume that economic agents use all the relevant information in forming their expectations of economic variables. The future is not fully predictable, but it is argued that economic agents' expectations are correct on average. In terms of the expected exchange rate, rational expectations imply that the future exchange rate equals the value expected for the currency at time $t$, given all the information at that time, plus an error term that is uncorrelated with that information. As a result, most tests of UIP are joint tests of UIP and the rational expectations hypothesis. Most of the literature has failed to find evidence of UIP.

## Purchasing Power Parity

*The law of one price* states that if markets are efficient—that is, markets reflect all relevant information—all identical goods must have only one price. This law has several implications for international transactions. For instance, two identical tradable goods, with no obstacles to international trade and no transactions costs, should have the same price (in the same currency) in two countries. That implies that the price of good $i$ in one country must equal its price in the foreign country multiplied by the exchange rate. This relationship can be represented by Equation 5:

$$S_t P_t^{i*} = P_t^i, \tag{5}$$

where $P_i^{i*}$ ( $P_t^i$ ) refers to the foreign (domestic) price of good $i$. The law of one price makes intuitive sense. If $P_t^i$ is greater than $S_t P_t^{i*}$, an investor will buy good $i$ in the foreign country, transport it to the domestic country, and sell it for a profit. But with time, these two prices will converge, until it is not possible to take advantage of this arbitrage opportunity. That is, the equality in Equation 5 will be restored once again.

Based on the law of one price, we can define the theory of purchasing power parity or PPP. While the law of one price is defined in terms of a certain good, PPP applies to the general

price level. PPP states that the exchange rate between two countries' currencies equals the ratio of the countries' price levels. This concept is represented by Equation 6:

$$S_t P_t^* = P_t$$

or also

$$S_t = \frac{P_t}{P_t^*}, \tag{6}$$

where $P_t^*$ $(P_t)$ is the foreign (domestic) price of a reference commodity basket. Therefore, a relative increase in the domestic price level will be associated with a proportional depreciation of the domestic currency. In this case, the domestic currency is losing relative purchasing power, and its value decreases. Note that in Equation 6, we do not need to use the subscript *i*, given that we are referring to the general price level, not the price of specific goods.

We can rearrange Equation 6 as

$$s_t = p_t - p_t^*, \tag{7}$$

where $p_t$ $(p_t^*)$ is the logarithm of $P_t$ $(P_t^*)$, and $s_t$ is the logarithm of $S_t$. In Equation 7, the logarithm of the exchange rate is simply the difference between the logarithms of the two prices. PPP does not hold empirically, among other things, because of transaction costs and obstacles to international trade. Therefore, economists have decided to call this version of PPP *absolute PPP,* while a more flexible version of PPP is called *relative PPP*. This relative version of PPP can be represented by Equation 8:

$$\Delta s_t = \Delta p_t - \Delta p_t^*, \tag{8}$$

where $\Delta s_t$ represents the change in $s_t$, and $\Delta p_t$ $(\Delta p_t^*)$ represents the change in $p_t$ $(p_t^*)$. Relative PPP states that the percentage change in the exchange rate between two currencies equals the difference between the percentage changes in the domestic and foreign price levels. If inflationary pressures force prices higher in one country but not another country, the exchange rate will change to reflect the change in the relative purchasing power of the two currencies.

The empirical evidence on PPP is mixed. Several empirical studies have failed to find evidence of PPP, while others have found some evidence of PPP in the long run. Economist and Nobel laureate Paul Krugman (1978) concludes in one his studies that deviations from PPP are large and fairly persistent.

## The Monetary Approach to the Exchange Rate

The monetary approach to the exchange rate is a theory of exchange rate determination in which exchange rates depend on the monetary aspects of the economy, that is, on money supply and money demand. To develop the predictions of this model, let us start by assuming that we have two countries in which money markets are equilibrated (i.e., money demand is equal to money supply). The money market in both countries can therefore be described as

$$m_t - p_t = a_1 y_t - a_2 i_t, \tag{9}$$

$$m_t^* - p_t^* = a_1 y_t^* - a_2 i_t^*, \tag{10}$$

where $m_t$ $(m_t^*)$ refers to the logarithm of domestic (foreign) money, $y_t$ $(y_t^*)$ is the logarithm of domestic (foreign) income, and $p_t$ $(p_t^*)$ and $i_t$ $(i_t^*)$ maintain the previous definitions as the logarithm of the domestic (foreign) prices and the domestic (foreign) interest rate, respectively. For both cases, an increase in interest rates—that is, an increase in the opportunity cost of holding money—causes the demand for money to fall. On the other hand, an increase in income, which encourages consumer spending, results in an increase in the demand for money for transaction purposes.

Equations 9 and 10 can be rearranged as

$$p_t = m_t - a_1 y_t + a_2 i_t, \tag{11}$$

$$p_t^* = m_t^* - a_1 y_t^* + a_2 i_t^*. \tag{12}$$

Solving for the relative price difference $(p_\tau - p_\tau^*)$, we get

$$p_t - p_t^* = (m_t - m_t^*) - a_1(y_t - y_t^*) + a_2(i_t - i_t^*). \tag{13}$$

Finally, substitute Equation 7 into Equation 13 to obtain

$$s_t = (m_t - m_t^*) - a_1(y_t - y_t^*) + a_2(i_t - i_t^*). \tag{14}$$

Therefore, if we assume that PPP holds, we get an equation that relates the exchange rate with money supply, income, and interest rates. An increase in the relative money supply $(m_t - m_t^*)$ leads to an increase in $s$, a depreciation of the domestic currency. Therefore, if a country increases its money supply faster than other countries, it will suffer a depreciation of its currency. Also, an increase in income leads to a domestic currency appreciation. Under this theory, an increase in income increases the transactions demand for money, and because there is a constant money supply, equilibrium in the money market can be achieved only if the domestic price falls. However, given that PPP holds, a drop in the domestic price is possible only if the domestic currency appreciates (see Equations 6 and 7). In a similar fashion, an increase in the domestic interest rate decreases money demand and depreciates the domestic currency.

**Table 42.1** Summary Predictions of the Monetary Approach to the Exchange Rate

| | *Increase in the . . .* | | |
|---|---|---|---|
| | *Relative Money Supply* ($m_t - m_t^*$) | *Relative Income* ($y_t - y_t^*$) | *Relative Interest Rate* ($i_t - i_t^*$) |
| Impact on the domestic currency | Depreciation | Appreciation | Depreciation |

Table 42.1 summarizes the predictions of the monetary approach to the exchange rate. As we mentioned above, an increase in the relative money supply and the relative interest rate will depreciate the domestic currency, while an increase in relative income will appreciate the currency. It is also possible to add expectations to the previous model. In a monetary model with rational expectations, the exchange rate depends not only on current excess money supply but also on expected future excess money supplies. In general, the effect of current relative changes in the money supply on the exchange rate depends on the perceived money supply rule. If the relative money supply increases but this change is expected to be just temporary, the expected future exchange rate would be only slightly affected, and hence the current exchange rate would simply reflect the current relative money supply change. If in contrast, the increase in the relative money supply leads to the expectation that domestic rates of monetary expansion would be higher than foreign rates in the future, then the domestic currency would depreciate by more than the current relative change in the money supply.

## Nominal and Real Exchange Rates

Until now, we have been referring to what economists call the nominal exchange rate. That is, we have been discussing the number of units of one currency that must be paid to acquire one unit of another currency. However, frequently individuals and companies are also interested in what can be bought with one unit of a currency in a certain country. Is it better to hold one U.S. dollar or one Japanese yen? That may depend on what can be acquired in the United States with a U.S. dollar and what can be acquired in Japan with one yen.

The *real exchange rate* is a measure of the real value of one unit of one currency with respect to one unit of the other currency. The real exchange rate is the product of the nominal exchange rate and the ratio of prices between the two countries. If absolute PPP holds, the real exchange rate is equal to 1. We can confirm this by rearranging Equation 6 to obtain

$$S_t\left(\frac{P_t^*}{P_t}\right) = 1. \tag{15}$$

A typical example of the real exchange rate is the relative price of a Big Mac in two countries. For simplicity, assume that Big Macs are the only goods in Japan and the United States. If the price of a Big Mac is $3 in the United States and ¥1 in Japan, while the nominal exchange rate is 0.0094 U.S. dollars for 1 yen, then the real exchange rate is

$$.0094\left(\frac{1}{3}\right) = .3333.$$

The real exchange rate is .3333, which indicates that Big Macs are relatively cheaper in the United States than in Japan. Therefore, it will make economic sense to acquire U.S. dollars and buy Big Macs in the United States, send them to Japan, and sell them for a profit there (assuming that Big Macs can survive the trip overseas!). Of course, any such opportunity for arbitrage will be short-lived; the additional demand for U.S. dollars is going to appreciate the U.S. dollar, and the real exchange rate will eventually converge to 1. However, as mentioned above, PPP does not hold in the real world because of transaction costs and obstacles to international trade. Therefore, the real exchange rate may deviate permanently from the value of 1. Still, the concept of the real exchange rate can be quite useful for determining whether a currency is overvalued or undervalued.

In the real world, it is also the case that there are many goods in a country, and therefore instead of using the price of just one good (e.g., Big Mac), economists use price indices from both countries to construct the real exchange rate. It is also the case that countries have more than one trading partner. Therefore, it is important to look at more than one bilateral real exchange rate for each country. A useful measure in this regard is the *real effective exchange rate* (REER). The REER is the average of the bilateral real exchange rates between one country and its trading partners, weighted by the respective trade shares of each partner.

## Conclusion

This chapter discussed the exchange rate, that is, the value of one currency in terms of another currency. Exchange rates are essential for economic transactions across national borders and are the principal component of the foreign exchange market. In this chapter, you also learned about the history of exchange rate regimes going from the pegged exchange rates, used during the gold standard era, to the mostly flexible exchange rates, which is what we have today.

Also discussed were the main theories of exchange rate determination, and in the process we familiarized ourselves with important concepts of the foreign exchange market, such as the difference between the spot exchange rate and the forward exchange rate. We were also able to discuss different theories related to these concepts, such as covered and uncovered interest rate parity and purchasing power parity.

As mentioned above, the specific factors that determine the value of a currency in terms of another currency remain the source of much debate. However, as we did in this chapter by looking at different theories of exchange rate determination, we can get a feel for the functioning of the exchange rate market, and we can relate the value of a country's currency with other important variables from that country, such as imports, exports, interest rates, income, and money supply. The next time that you travel abroad, purchase a foreign good, or get involved in any type of international transaction, take a minute to analyze how the theories discussed in this chapter apply to your transaction. You may even come up with new ideas to formulate better theories of exchange rate determination.

## References and Further Readings


Bank for International Settlements. (2002, March). *Triennial Central Bank Survey: Foreign Exchange and Derivatives Market Activity in 2001—Final Results.* Retrieved August 10, 2008, from http://www.bis.org/publ/rpfx02.htm

Bordo, M. (1981, May). The classical gold standard: Some lessons for today. *Review of the Federal Reserve Bank of St. Louis,* pp. 2–17.

Bordo, M. (1993, March). The gold standard, Bretton Woods and other monetary regimes: A historical appraisal. *Review of the Federal Reserve Bank of St. Louis,* pp. 123–191.

Bordo, M., & Schwartz, A. (1984). *A retrospective on the classical gold standard, 1821–1931.* Chicago: University of Chicago Press for the National Bureau of Economic Research.

Calvo, G., & Reinhart, C. (2000). *Fear of floating* (National Bureau of Economic Research Working Paper Series 7993). Cambridge, MA: National Bureau of Economic Research.

Cross, S. (1998). *All About . . . the foreign exchange market in the United States.* Retrieved August 1, 2008, from http://www.newyorkfed.org/education/addpub/usfxm

Edwards, S. (1989). *Real exchange rates, devaluation, and adjustment exchange rate policy in developing countries.* Cambridge: MIT Press.

Eichengreen, B. (1998). *Globalizing capital: A history of the international monetary system.* Princeton, NJ: Princeton University Press.

Fama, E. F. (1984). Forward and spot exchange rates. *Journal of Monetary Economics, 14,* 319–338.

Fischer, S. (2001). Exchange rate regimes: Is the bipolar view correct? *Journal of Economic Perspectives, 15,* 3–24.

Frankel, J. (1999). *No single currency regime is right for all countries or at all times* (National Bureau of Economic Research Working Paper Series 7338). Cambridge, MA: National Bureau of Economic Research.

Frenkel, J. A. (1978). Purchasing power parity: Doctrinal perspectives and evidence from the 1920s. *Journal of International Economics, 8,* 169–191.

Frenkel, J. A., & Levich, R. M. (1975). Covered interest rate parity: Unexploited profits? *Journal of Political Economy, 89,* 1209–1224.

Friedman, M. (1953). The case for flexible exchange rates. In *Essays in positive economics* (pp. 157–203). Chicago: University of Chicago Press.

International Monetary Fund. (2004). *Classification of exchange rate arrangements and monetary policy frameworks: What the IMF does.* Retrieved August 2, 2008, from http://www.imf.org/external/np/mfd/er/2004/eng/0604.htm

Johnson, H. (1969, June). The case for flexible exchange rates, 1969. *Review of the Federal Reserve Bank of St. Louis,* pp. 12–24.

Krugman, P. (1978). Purchasing power parity and exchange rate: Another look at the evidence. *Journal of International Economics, 8,* 397–407.

Obstfeld, M., & Rogoff, K. (1995). *The mirage of fixed exchange rates* (National Bureau of Economic Research Working Paper Series 5191). Cambridge, MA: National Bureau of Economic Research.

Mundell, R. (1961). A theory of optimum currency areas. *American Economic Review, 51,* 509–517.

Tsiang, S. C. (1959). The theory of forward exchange and effects of government intervention in the forward exchange market. *IMF Staff Papers, 7,* 75–106.

# 43

# Comparative Economic Systems

Satyananda J. Gabriel
*Mount Holyoke College*

Comparative economic systems, as a distinct subdiscipline of economics, began in the United States during the 1930s with the publication of textbooks such as William N. Loucks and J. Weldon Hoot's (1938) *Comparative Economic Systems.* However, as Maurice Dobb (1949) would note in his critical review of Loucks's text on the subject, comparative economic systems was already a thriving field of study in Europe, especially Britain. Benjamin Ward (1980) argues that comparative economic systems emerged as a field of study initially because of the utopian novels of the later nineteenth century and practicing communes whose existence both influenced the novels and was influenced by them. Novels such as Edward Bellamy's (1888) *Looking Backward* helped to spur political mass movements and inspired the development of utopian communities, perceived as alternative economic and social systems to those prevalent at the time.

Undoubtedly this was in reaction to the economic and political upheavals that had come with the Industrial Revolution and its aftermath. This was a time of epochal change, as the emergence of capitalism, colonialism, and the introduction of new technologies and new relations of production caused disruptions in traditional ways of life. In the trauma of massive social change, displacement of rural and urban artisans, and radical shifts in income distribution and political power, large numbers of people came to question dominant theories of economic organization and distribution. The notion of comparing alternative economic systems was a direct reaction to witnessing the emergence of a new economic and social system and the disintegration of an older one.

Intellectuals during this period investigated what economic system would work best for the productive organization of a nation. This became an even more urgent matter from late 1929 through the 1930s, when many nations suffered through a global economic depression that caused many to question whether capitalism was really the road to wealth rather than an engine of wealth destruction. However, the methods used by social scientists in that early period were varied, as were the definitions and typologies deployed in comparative economic systems analysis. The subdiscipline did not gain a reasonably coherent set of definitions and typologies until the 1950s. It was during this period that the influences of the cold war were felt most strongly and shaped the contours of comparative economic systems analysis and teaching. Competition between the United States and the USSR, which had been first in space with *Sputnik,* stimulated efforts at careful study of these two alternative economic systems.

Substantial support from government and foundations was forthcoming for the study of a topic that now was of central policy relevance. Research institutes were established, especially at Harvard University and Columbia University, to produce research that provided a firm basis for appraisal of these competing economic systems. Economists such as Abram Bergson, Lenny Kirsh, Leon Smolinski, Joe Berliner, Frank Holzman, Barney Schwalberg, Marshall Goldman, and Padma Desai at The Russian Research Center at Harvard University became pivotal figures in shaping the direction of research in the field and cementing a dualistic approach of analysis pitting "East" versus "West."

In the early 1960s, as America's political leadership began to seek solutions to a perceived gap in the rate of scientific achievement between the USSR and the United States, comparative economic systems analysis would turn markedly toward quantitative tools for evaluating these competing economic systems. Abram Bergson's (1961) text *The Real National Income of Soviet Russia Since 1928* became a seminal text. Gur Ofer (2005) argues in a biographical essay on Bergson that it was Bergson who shaped the dominant methodology for estimating Soviet performance in gross national product (GNP) statistics: measuring the structure and rate of growth of the USSR in these terms and using these numbers as the starting point for comparison of the USSR and the United States. Bergson relied on his statistical analysis to reject the merits of "socialism," as represented by the USSR.

This prompted an ongoing debate over methodological issues. Many in the field, such as Philip Hanson (1971) tried to show that "efficiency" comparisons using standard quantitative techniques could not provide an objective assessment of the comparative efficiency of U.S. versus USSR economic institutions. As Hanson would argue, "The chief difficulty in measuring the comparative performance of different economic systems is an obvious but fundamental one. The problem is to isolate the system and its effects from other influences which lead to measurable international differences. Or, to put it another way, the problem is to identify the system and isolate its performance from the environment in which the system has had to perform" (p. 327). In other words, the relative performances of the United States and the USSR were sensitive to the initial conditions predating their respective contemporary economic systems. A country's history matters.

Comparativists continued to produce varied analyses and, more important, varied models of "systems," both quantitative and qualitative, but the cold war orthodoxy came to dominate the content of undergraduate textbooks and courses in comparative economic systems. Capitalism (or "free markets") was contrasted with socialism (or "communism"), where the features of the former were gleaned from aspects of the United States and its North Atlantic Treaty Organization (NATO) allies, and the latter was created from idealized features of post-Lenin USSR and its Council for Mutual Economic Assistance (1949–1991) allies. Instances of free markets within the latter grouping, such as markets in beyond-quota agricultural goods and consumer goods, and instances of state planning, such as by central banks and various ministries or agencies within the former grouping, were ignored in favor of the stylized market versus command bipolarity. The earlier lack of consensus on the appropriate typology to use in comparative economic systems analysis gave way to this bipolar approach, even though the definitions employed for the two primary systems could be radically different (or, in many cases, so poorly articulated that readers were left to presume the meaning of the terms from prior readings or, more likely, popular discourse). *Capitalism, socialism, market socialism, communism, self-employment, petty commodity production, peasantry, feudalism, free markets, command economies, centrally planned economies, indicatively planned economies,* and so on were all terms disseminated in texts and journal articles, often with the same term having different meanings, depending on the authors.

The end of the cold war was the catalyst for a return to the uncertain foundations reminiscent of the 1930s. Many comparativists morphed into "transition" economists, and many of the courses in the field have been similarly transformed into courses on economic transition. Paradoxically, these courses remain essentially rooted in the same polemics of the 1950s cold war paradigm, with the explicit assumption that idealized market structures, institutions, and processes represent the Hegelian endpoint of economic progress. The questions related to comparing alternative economic systems and their relative merit had completely given way, in transitions studies, to examining the various paths of former socialist nations to free-market capitalism. Thus, in an odd twist, this transition studies version of comparative economic systems represents a convergence with the teleological position of orthodox Marxian theory, albeit with a different presumed telos (the market economy is represented as the end of history, whereas Marxism viewed communism as the end of history).

The spin-off of transition studies from comparative economic systems may ultimately be beneficial to the subdiscipline. Comparativists, freed from cold war rhetoric, have already begun to explore the broader terrain indicated by Hanson (1971) and others. Andrew Zimbalist (1984) has written that "the scope of comparative economic systems as a field singularly offers the potential, inter alia: (a) to explore and challenge the assumptions and methods of traditional economic analysis; (b) to reinterpret conventional wisdom; (c) to understand the interplay of economic and noneconomic forces in different institutional contexts; and (d) to evaluate the desirability of alternative economic policies and structures" (p. 1).

The study of comparative economic systems is, arguably, the subdiscipline of economics that fosters the broadest range of methodologies. Comparativists take as their field of study the microeconomic and macroeconomic relationships in the economy, institutional structures, political and cultural processes, and even demographic and geographic differences between social formations. The tools used in research are, therefore, quite varied. The study and analysis of comparative economic systems gives more weight to political institutions as determinants of economic structures and outcomes than any other subdiscipline of economics, except for public economics. Many comparativists focus their research primarily on microeconomic relationships, such as market structures and pricing mechanisms. Theoretical models, case studies, and empirical analyses are all important tools in

comparative economic systems analysis. The unique perspective of comparative economic systems analysis can be seen in a variety of contexts in the history of economic thought.

## The Socialist Controversy

Perhaps one of the most interesting and confusing juxtapositions in comparative economic systems is that between socialism and capitalism. There has been a long-running debate among comparativists not only about which of these economic systems is better for a society but also what constitutes "socialism" and "capitalism."

The debate over the relative economic viability of "socialism" versus "capitalism" began in the formative years of comparative economic systems. It appears to have been sparked by Ludwig von Mises's (1920) "Economic Calculation on the Socialist Commonwealth," a scathing attack on the viability of socialism as a legitimate economic system:

> Economics, as such, figures all too sparsely in the glamorous pictures painted by the Utopians. They invariably explain how, in the cloud-cuckoo lands of their fancy, roast pigeons will in some way fly into the mouths of the comrades, but they omit to show how this miracle is to take place. Where they do in fact commence to be more explicit in the domain of economics, they soon find themselves at a loss—one remembers, for instance, Proudhon's fantastic dreams of an "exchange bank"—so that it is not difficult to point out their logical fallacies. (p. 2)

Von Mises's (1920) main argument was that if the state owned the factors of production, then there would be no market for capital goods and therefore no pricing mechanism to calculate a rational value for such goods. In the absence of a rationally determined value of capital goods, there can be no rational allocation of capital or consumer goods in the society. Defining socialism in terms of this sort of state control over investment and allocation, von Mises argued against direct government involvement in the economy, in general, and against the idea of central planning, in particular. Planners, faced with an enormously complex operations management and product allocation puzzle, would have no idea about the relative feasibility or value-generating (or destroying) potential of various options for investment. Von Mises's polemical argument against the legitimacy of socialism would become a key tenet of the so-called Austrian and Chicago schools of economics and would incense those who supported socialism. His work ignited a debate that continues to the present day.

In a series of articles titled "On the Economic Theory of Socialism," Oskar Lange (1936) presented the best-known rebuttal to von Mises on the economic aspects and organization of a socialist system. He argued that

> there is not the slightest reason why a trial and error procedure, similar to that in a competitive market, could not work in a socialist economy to determine the accounting prices of capital goods and of the productive resources in public ownership. Indeed, it seems that it would, or at least could, work much better in a socialist economy than it does in a competitive market. For the Central Planning Board has a much wider knowledge of what is going on in the whole economic system than any private entrepreneur can ever have; and, consequently, may be able to reach the right equilibrium prices by a much shorter series of successive trials than a competitive market actually does. (p. 67)

Both von Mises and Lange were Austrian economists, although Lange had no affinity for the Austrian school of economics. On the other hand, Lange did share von Mises's love of neoclassical pricing theory, which assumed that having large numbers of buyers and sellers was the most efficient method of generating useful economic information. Despite his belief in the utility of this pricing dynamic, Lange advanced the proposition that capitalism was prone to breaking down and creating crises, resulting in the waste of resources and human potential. Principal among these malfunctions was in the optimal distribution of incomes and maximization of social welfare in the society. In addition, Lange argued that capitalism does not count all the costs of production, only those that directly affect the income of the capitalist enterprise. This argument has echoes in current debates over pollution, greenhouse gases, and other externalities.

Furthermore, Lange asserted that a socialist economy could avoid the bogeyman of capitalism—the business cycle—because of the state's ability to permit lots of small-scale markets with many buyers and sellers while simultaneously using the planning bureaucracy to make adjustments to capital investment so as to avoid unemployment. The argument that government agencies could marshal the data necessary to manage the economy, even while economic units remained relatively decentralized, had a wide following in the 1930s and 1940s. The mathematical arguments and statistical works of Lange and perhaps even more so the mathematical models of Wassily Leontief contributed to an active discourse on the potential for government to manage an economy through the application of mathematical models.

Real-world examples of such attempts to manage economies never quite lived up to expectations. Arguably, the best real-world examples of Langean socialism, at least prior to the rise of post-Mao China, were Yugoslavia after 1965 and Hungary from implementation of the 1968 "new economic mechanism" to the end of the Council for Mutual Economic Assistance (CMEA) era, although neither Yugoslavia nor CMEA-era Hungary nor post-Mao China met Lange's requirement that the state bureaucracy be submitted to the democratic control of the citizenry. In the absence of such democratic means for disciplining the government, agency costs and poor planning choices could

occur with very little risk that state officials could be removed from power.

While Lange argued for a decentralized and market-based form of socialism, he described the actual capitalist economy (as opposed to the theoretical capitalist economic system in textbooks) as primarily dominated by oligopolistic and monopolistic firms, rather than powerless firms operating in a world of perfect competition. It is the latter (perfectly competitive) economic system that serves as the default within microeconomic textbooks and is a fundamental condition for laissez-faire (free-market/neoliberal) public policy prescriptions. For Lange, market power was a fact of economic life and led to suboptimal economic decisions within capitalism. Lange further raised the question of whether, under oligopolistic/monopolistic conditions, capitalism could serve as the social system best able to continue human economic and social progress. Thus, Lange was directly addressing one of the fundamental concerns that had given birth to comparative economic systems and throwing down the gauntlet to those who would support the notion that the prevalent capitalist economic system was the best of all possible worlds.

In the 1940s, Friedrich A. von Hayek, a peer and ally to von Mises, took up the challenge posed by Lange. Hayek rebutted Lange's premises and conclusions about the relative efficacy of socialism versus capitalism in a series of articles and books. Hayek challenged the rationale of state-directed pricing and resource allocation and advocated the necessity of a completely market-based system for any rational distribution of products in society. His most notable contributions to the debate were the idea of limits of human cognition and the theory of knowledge dispersion. Contrary to Lange's argument that a central planning bureaucracy could have much wider knowledge of an economy, Hayek argued that the central planning process was untenable because the information required to successfully manage an economy was beyond the comprehension of any bureaucratic structure. The planners within such a structure would suffer simultaneously from insufficient information and information overload. For Hayek, a planning bureaucracy could not replace thousands of entrepreneurs and consumers in generating and processing the information necessary for a successful economy. And because the planning bureaucracy could not possibly gain or process enough information to make good decisions, bureaucrats would tend to be risk averse and not make the sorts of investments in the economy necessary for growth and development. Absent the incentives of a market system, central planners and industrial managers would have no motivation to make good choices about the use of firm assets and generate what we today refer to as agency costs. Hayek argued that competitive markets generate knowledge discovery at lower costs and incentives for managers to make superior choices of asset deployment, production technologies, and investment. He even theorized that it was only through competitive markets that the question of goods and their qualities and quantities, even their existence as goods, could be known. He theorized that given the dispersion of existing knowledge of the needs and wants of economic actors, a centralized bureaucracy would not even know where to begin to make sense of the optimal allocation of goods and resources such that this allocation could be combined into a single plan. Hayek felt that the state under socialism had an impossible task because knowledge of preferences and other relevant economic data is localized, subjective, and difficult (to nearly impossible) to pass on to others in complete, undistorted, and useable forms. For Hayek, the market system and other supporting institutions were examples of *spontaneously organized complex phenomena* that were beneficial in the development of societies and individual initiative and achievement. The market could achieve what no state could achieve. The economic system promoted by Hayek and other members of the Austrian (and Chicago) schools was one where the state played a limited role, primarily in the provision of a minimal social insurance for citizens, the protection of individual property rights, and military defense of the nation as a whole.

One of the many critiques of Hayek and von Mises (and, in a larger sense, of the Austrian and Chicago schools) was their failure to recognize that both market capitalist systems, where private corporate structures prevail, and socialist systems, with their state bureaucracies, could suffer from agency costs, as well as incentive and allocation problems due to asymmetric information (Caldwell, 1997) and asymmetric power. In many ways, Lange's challenge to the notion of a free-market capitalist system, as imagined within the theoretical work of the Austrian and Chicago schools, is not really answered, whether or not his broader conclusions about socialism are accepted. In other words, Lange could be incorrect about the efficiency of socialism and yet correct in his conclusions about the inefficiencies of actual capitalism. The reason the challenge is not taken up by von Mises and others is that they have tended to ignore the features of actual economic systems in societies deemed to be capitalist in favor of imaginary economic systems of completely disaggregated and powerless economic agents: Even firms are conceptualized as simply the congealed sites of the actions of individual economic agents. If Lange is correct that markets are epitomized by various forms of market power, then there is no reason to assume optimal or even rational social outcomes. This conclusion would be even stronger in the presence of externalities, as well as agency costs and related information asymmetries. In other words, the challenge of comparing economic systems remains, despite the implicit (and sometimes explicit) assumption of the Austrian and Chicago schools that this is now a dead question.

However, the rise of the USSR after World War II and deterioration of relations with the United States made it difficult for any objective analyses of alternative economic systems as the socialism–capitalism debate increasingly gave

way to polemics. Comparative economic systems served as a sort of ground zero within economics for these polemics. The subdiscipline came to play an instrumental role in critiquing the Soviet bloc, more than a source of scientific discourse about systemic differences. Indeed, the fact that the world, present and past, provided a rich source of complex and different economic systems, even within countries described as capitalist, was lost in an increasingly dualistic argument about the merits of the West versus the East. The term *command economies* (first used by Grossman, 1963) was coined, although most discussions of the USSR and other CMEA nations used the term *communism* because most of the parties in power in that part of the world called themselves communist, placing their end goal (telos) at greater prominence than the adjective they used to describe their economic systems (socialist). Many of these countries were, however, labeled by their new leaders *people's democracies,* a term that should have presaged the development of extensive democratic institutions but instead highlights the degree to which words were being used polemically.

Comparative economic systems texts described an idealized set of dominant politico-economic institutions of the NATO countries (the "West") and a similarly idealized set of dominant politico-economic institutions of the CMEA countries (the "East") as *the* central duality among global economic systems. The basic dichotomy generated by this polemic was an outgrowth of the Austrian and Chicago schools: the dichotomy between "free-market" or "market capitalist" economies and "command" or "communist" economies. A few texts continued to use the proper term *socialist* for the "people's democracies" of the Eastern bloc. Some comparativists used terms such as *command socialist* or *centrally planned* economies. In any event, in most discussions and textual materials, the distinction between the two poles of this duality was meant to be absolute: thesis counterposed to antithesis. While it is undeniable that the experiment in centralized, command planning in the USSR and the larger CMEA was both noteworthy and distinct from the more decentralized planning in U.S.-style capitalism, it is a stretch to go from this particular distinction to the idea that these two examples of social formations had nothing in common (both in terms of social institutions and processes). Command elements in the NATO politico-economic social structures were largely or completely ignored, as were similarities in the underlying economic relationships within NATO and CMEA firms (including both command and central planning aspects, even if in microcosm). This dichotomy further ignores the flexibility of capitalism and the constantly changing mix of state intervention (into market exchange relationships, the conditions shaping worker freedoms or lack thereof, firm governance, etc.), including occasional direct state involvement in productive investment and production even in the most liberal NATO member states, such as the United States, during historical periods when "free-market" approaches were deemed insufficient to resolve economic crises.

The nature of socialist economic systems has similarly been protean in character. In the non-Marxian literature, socialism is rarely delineated in a clear manner. A wide range of economic systems have been labeled socialist (Schnitzer, 2000). Socialism is sometimes said to require the replacement of private property in the "major means of production" with public (state) ownership. Alternatively, "market" mechanisms are recognized as compatible, if not integral, to socialist economic systems. Indeed, the term *market socialism* has long been part of the literature and was the standard designation for the economic system of post-1965 Yugoslavia. Government planning to achieve economic and social objectives such as redistribution of income through progressive taxation, transfer payments, and provision of public goods has often been an element in a subset of social scientific and most polemical definitions of socialism. However, it would be difficult to find a genuinely viable state where there was not some degree of government planning, income redistribution, provision of public goods, and so on. A few comparativists have defined socialism as an economic system within which democratic decision making has been extended to the economic sphere (Zimbalist, Sherman, & Brown, 1989). In this vein, there are those who discuss "worker-managed market socialism," highlighting the issue of who controls the firm. Many of the comparativists who studied the former Yugoslavia believed the worker-managed market socialism of that country was an important *socialist* alternative to the Stalinist command socialism of the USSR. And there are social scientists who view worker-managed firms in the United States as a form of proto-socialism. These theorists often have a very different view of the role of the state under socialism, arguing for the viability of a decentralized form of socialism that does not require an extensive state bureaucracy or central planning. Neoclassical economists have obscured the term even further by defining *welfare state* economies based on some of the same characteristics others have associated with socialism. Orthodox Marxist theorists have added to the confusion and ambiguity over the term *socialism.* To further complicate any attempt at consistent typology, Marxian theory has its own longstanding definition of socialism. Paul Sweezy (1976) put it quite succinctly when he defined socialism as "a way station between capitalism and communism."

This "blurring" of characteristics among economic systems and the ambiguity of terms such as *capitalism* and *socialism* plague many theoretic approaches to comparative economic systems. Comparative economic systems depends critically on the typology deployed in analyzing distinct economic systems, understood to be alternatives and evaluated on the basis of scientific methodology. The bipolar capitalism/socialism debate has been mostly polemical, although important questions arose within that debate and have inspired analysis, both empirical and theoretical, from contemporary comparativists. Thus, the monochromatic

understanding of economic systems that came out of the cold war debates has gradually begun to give way to a more complex set of analyses. These studies are contributing to a better understanding of the role of different institutions and alternative economic processes. Unfortunately, the polemics of the cold war tainted discussions of future economic systems. After all, communism referred specifically to an economic system that had not yet come into being but served as the ideal (telos) around which orthodox Marxists constructed their arguments in favor of revolutionary change in societies. Socialism, in the traditional Marxian framework, is understood as the transitional stage (where it is possible for capitalism to continue to predominate in the economic realm) on the road to that ideal society. Thus, to the extent this logic of transition to an ideal society came to be associated with the Stalinist political apparatus and the struggle between West and East, the very notion of speculating about future economic systems came to be discredited within mainstream economics. Perhaps eventually the interest in comparative economic systems as a way of thinking about future societies and economies will return to the fore after the lingering bad taste of the cold war is forgotten. Just as the imaginations of mechanical engineers dreaming of building better machines can stimulate invention and innovation, perhaps economists would be better at their craft if they could dream of better societies.

## Class-Based Typologies

An alternative to the old bipolar approach to systems typologies is one based on the microeconomics of class. This approach has been used to construct a typology of alternative economic systems (or modes of production) on the basis of systems for organizing labor and distributing the fruits of that labor.

The term *capitalism* has now been part of social discourse for such a long time that it is easy to forget that the term has its origins in the writings of Marx. It is worth noting that Marx's critique of political economy and analysis of the capitalism of the early nineteenth century came from his use of a framework that compared alternative economic systems. In other words, Marx was probably the first genuine comparativist.

Class-based typologies of economic systems come from definitions invented by Marx to describe a variety of social systems but primarily to understand capitalism, in part, by contrasting it to these other systems. This approach focuses attention on the relationship between direct producers (workers whose productivity results directly in the creation of new products) and alternative social mechanisms by which control over the distribution of economic profits is exercised. While many Marxian and non-Marxian social scientists have conflated ownership systems with control over profit distributions, this has not been the case with poststructuralist Marxian theorists, such as Stephen A. Resnick and Richard D. Wolff, authors of *Knowledge and Class* (1987), who have written extensively on the application of class concepts to understanding alternative economic systems. In the Resnick and Wolff framework, class is defined by the unique social system by which the surplus labor/product of direct producers is created, appropriated, and distributed.

In this paradigm, capitalism is defined by the unique manner in which surplus labor is generated from free wage laborers who do not control the receipt and distribution of the fruits of that labor (in the form of product or money). Their definition allows for a very wide range of variant forms of capitalist societies. If the type of capitalism is defined by who appropriates and distributes the economic profits or surplus, so long as the profits come about because of the work of free wage laborers, it becomes possible to speak of state capitalism, where a state bureaucracy "exploits" workers, or corporate capitalism, where the appropriating/distributing institution is a private corporation, or even church capitalism, if the "capitalist" is a religious institution. It becomes possible to investigate, both theoretically and empirically, the implications of these alternative forms of capitalism as part of a comparative economic systems research agenda.

## Transition Economics

The notion of transition from one economic system to another can be explained in evolutionary terms, as either an accident or movements along a teleological path. The early discussion of transitional economic systems was centered on the transition to socialism. Dorothy Douglas's (1953) influential text, *Transitional Economic Systems,* came out of this tradition and examined the difficulties of transition to socialism for Poland and Czechoslovakia. Thus, the notion of socialism as a mature form of capitalism is completely consistent with not only the Marxian teleology but also more contemporary teleological thinking, including the notion of some type of combined liberal democracy and capitalism as a systemic "end of history," in which socialism, rather than a higher stage (as in the Marxian conception), is perceived as a lower stage. In this inversion of the Marxian dialectic within a subset of comparative economic systems discourse, it is understood that the dynamic by which transition occurs is one in which societies select for liberal democracy and capitalism (and the usual assumption that the former political system is necessarily wedded to the latter economic system) after the failure of previous economic systems, including the socialist brand. The timing of this selection process will be different, and some societies will languish because of a breakdown in or structural impediments to the transition process.

Consider the case of the USSR, which took more than 75 years to transition out of "socialism." Analysts can point to specific structural impediments to transition, such as police state tactics, dependence on the central planning

mechanism on the part of many managers within the state bureaucracy, and so on. The usual presumption is that this transition out of socialism constituted a radical transformation not only in all aspects of the political structure but in all the economic processes as well. However, recall that the Marxian notion of socialism does not require the absence of a capitalist economy. Indeed, it was commonplace to argue that socialism was a transition in political authority (from a "dictatorship of the bourgeoisie" to a "dictatorship of the proletariat") but with a still extant capitalist economy. Resnick and Wolff (2002) recently have described the USSR as state capitalism and produce a completely plausible argument based on a poststructuralist Marxian framework. This approach has very serious implications for a post–cold war rethinking of comparative economic systems. It may be of great utility to differentiate among variant forms of capitalist systems, as with socialist systems or feudal systems in the contemporary world.

## Comparative Economic Systems and the China Puzzle

China has always been an important social formation within the comparative economic systems literature. Mao's deviation from the Leninist-Stalinist orthodoxy established by the USSR provided comparativists a major alternative version of socialism. And given the turmoil of the Maoist period, this alternative was almost always in flux, undergoing major structural and process changes. There was never a dull moment for the comparativist who chose to study China.

Nevertheless, Chinese socialism did have features in common with the USSR and CMEA bloc. The Communist Party–led bureaucracy suffered from poor incentives, high agency costs, resource bottlenecks and distribution problems, and serious constraints upon and therefore a shortage of entrepreneurial creativity of the sort that Joseph Schumpeter believed necessary for long-term economic growth and development.

The Maoist period saw the creation of communes that, in fact, resembled feudal domains in that the Communist Party hierarchy maintained tight control over value generated within the communes and bound farmers and their families to these land-based structures: Internal migration was strictly controlled by a household registration system. Most urban workers were tied just as tightly to state-owned enterprises that would typically include schools for workers' children, hospitals and clinics for workers' health care, and institutions providing other vital services. The link between the state and workers and farmers in China was far more all-inclusive and difficult to break than was the case in the Stalinist world of the USSR. Despite Mao's often radically democratic rhetoric about granting workers and peasants a primary decision-making role in the society, the direction of authority ran almost exclusively from the top (state) to the bottom (workers). Nevertheless, the gradual transformation of this system into a more Langean version of decentralized socialism, with a powerful state bureaucracy and planning institutions intact, would begin a mere 2 years after Mao's death in 1976.

The result of this transition is the creation of a form of market socialism that has been extraordinarily successful. Perhaps Lange has finally been exonerated. If we examine the key processes that have been the focus of comparative economic systems—product circulation and capital budgeting processes (market vs. planning), ownership structures (private vs. state), and political structures (liberal democratic states vs. single-party totalitarian states)—it is apparent that the neoclassical/neoliberal orthodoxy, an outgrowth of the Austrian and Chicago schools' dominance of economic theory from the 1950s onward, has problems with China's success story. This success story does not map neatly onto the usual grid of presumed one-to-one correspondences between systemic structures and economic outcomes. China's phenomenal growth rates and development have come about despite its deviation from orthodox policy prescriptions and continued deviation from the form of liberal democratic state that has been presumed to be a condition for sustainable economic growth by comparativists and others within mainstream economics and political science. In the words of Gary Jefferson (2008), "China's experience offers a now irrefutable lesson. Economic transition and development is a far more complicated phenomenon than simply putting in place the principal elements of the neoclassical model, whose policy implications have been dubbed the Washington Consensus, or even adhering to an enlarged doctrine that incorporates the basic tenets of the New Institutional Economics" (p. 170).

Established orthodoxy in comparative economic systems asserts that capitalism, especially the "free-market" variation, is the superior economic system. Policy makers and advisers have pushed this prescription for economies in transition with varying results. China's deviation from this prescription has renewed debate in the field about the relevance and accuracy of various theories as applied in actual practice. For instance, it is accepted that China has experienced rapid economic growth in the context of weak and often ambiguous property rights; an authoritarian political system; high agency costs from corruption within various locations in the state bureaucracy, which include state-owned enterprises; low levels of transparency in government and enterprise organizations; and a weak legal and law enforcement system. Yet, the Chinese economy has outperformed the United States and most other capitalist economies. Advocates of the "shock therapy" approach to economic transition in which rapid market liberalization and structural reform, if not dismantlement, of government bureaucratic structures have been humbled by the relative performance of economies that followed their prescriptions versus nations, particularly China, that have taken radically different paths to transition from one economic system to another. China's "gradual" reforms have been a success story of enormous proportion. China allowed public ownership to play a major

role in the economy well into its years of economic reform, and it remains a critical cornerstone in its transitional economic plan, especially in terms of the financial system, but also in the state-owned enterprise sector. Chinese authorities have simultaneously tapped the equity markets and maintained controlling interest in a large number of industrial conglomerates that were taken public. In other words, it would seem that China is following a path that has some elements in common with the arguments of Oskar Lange, who believed it possible to have a socialism that relied heavily on the market yet maintained considerable government involvement in and management of the economy.

The Chinese strategy flies in the face of an economic orthodoxy that, in keeping with Austrian and Chicago school dogma, posits economic freedom and individualism as *essential* preconditions for growth. The notion that "socialism" is an inferior economic system vis-à-vis "capitalism" has been severely threatened by the rise of China. It is increasingly common for policy makers, intellectuals, and others in the lower income nations to turn to China as the model economic system and source of the secret formula for their own nations to rise.

## Comparative Economic Systems in the Twenty-First Century

In charting a new direction for the field, many comparativists accept the notion that variant forms of capitalism predominate in the global landscape. Djankov, Glaeser, La Porta, Lopez-de-Silanes, and Shleifer, among others, argue that the new comparative economics should focus on these alternative capitalist models. However, rather than following the old systems-as-a-totality approach, these comparativists appear to favor an approach closer to that of the poststructuralist Marxian theorists, who prefer to isolate specific institutions or processes as the source of variation. The implicit institutionalism gaining currency in the subdiscipline is coupled with increasing use of econometric techniques in comparative analysis. Comparativists such as Montias, Ben-Ner, and Neuberger (1994) argue that if comparative analysis is to be useful for policy purposes, individual effects of economic and noneconomic variables must be identified. As John Bonin (1998) argues,

> Country-specific conditions clearly affect outcomes and the choice of policies depends on the system norms. Furthermore, the preferences of the economic actors themselves are influenced by interactions with other systemic components so that some environmental characteristics can be considered endogenous. This rich complexity of interdependencies dictates a need for careful recognition of problems of endogeneity and skillful use of modern analysis to isolate crucial relationships. (p. 5)

Montias and others (1994) argue that comparative systems analysis should discard the outdated practice of defining "systems" and embrace the new work of defining *institutions.*

Of course, there is dissension among the ranks. There are some such as Geert Hofstede, Amir Licht, Chanan Goldschmidt, Shalom Schwartz, and Frederic Pryor who argue that it is culture that should be the primary object of analysis. These comparativists start from the premise that cultural values determine not only why specific economic systems are adopted in different countries but also how these systems come to exist in the first instance.

It is clear that the study of comparative economic systems, like the rest of economics, faces new challenges in a twenty-first century where economic and social transformation appears to be speeding up. There should be no anticipation of finding any singular paradigm that satisfies all possible research questions. Institutional and neoclassical economists, the German-historical school, poststructuralist Marxian theorists, post-Hegelians, and others have and will continue to tackle the questions of typology of social formations and the issue of transition, positing a wide range of concepts relevant to such analyses.

## References and Further Readings


Bellamy, E. (1888). *Looking backward.* London: William Reeves.

Bergson, A. (1961). *The real national income of Soviet Russia since 1928.* Cambridge, MA: Harvard University Press.

Berliner, J. (1964). The static efficiency of the Soviet economy. *American Economic Review, 54,* 480–489.

Bonin, J. (1998). The "transition" in comparative economics. *Journal of Comparative Economics, 26,* 1–8.

Caldwell, B. (1997). Hayek and socialism. *Journal of Economic Literature, 35,* 1856–1890.

Desai, P. (1989). *The Soviet economy.* London: Blackwell.

Djankov, S., Glaeser, E., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2003). The new comparative economics. *Journal of Comparative Economics, 31,* 595–619.

Dobb, M. (1949). Review: *Comparative economic systems: Capitalism, socialism, communism, fascism, co-operation* by W. N. Loucks and J. W. Hoot, New York: Harper Brothers. *The Economic Journal, 59,* 426–429.

Douglas, D. (1953). *Transitional economic systems.* London: Routledge & Kegan Paul.

Ericson, R. (2006). Command versus "shadow": The conflicted soul of the Soviet economy. *Comparative Economic Studies, 48,* 50–76.

Goldman, M. (1968). *The Soviet economy: Myth and reality.* Englewood Cliffs, NJ: Prentice Hall.

Grossman, G. (1963). Notes for a theory of the command economy. *Soviet Studies, 15,* 101–123.

Hanson, P. (1971). East-West comparisons and comparative economic systems. *Soviet Studies, 22,* 327–343.

Hayek, F. A. von. (1944). *The road to serfdom.* Chicago: University of Chicago Press.

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

Jefferson, G. (2008). How has China's economic emergence contributed to the field of economics? *Comparative Economic Studies, 50,* 167–209.

Lange, O. (1936). On the economic theory of socialism: Part one. *Review of Economic Studies, 4,* 53–71.

Lange, O. (1937). On the economic theory of socialism: Part two. *Review of Economic Studies, 4,* 123–142.

Licht, A., Goldschmidt, C., & Schwartz, S. H. (2007). Culture rules: The foundations of the rule of law and other norms of governance. *Journal of Comparative Economics, 35,* 659–688.

Loucks, W. N., & Hoot, J. W. (1938). *Comparative economic systems.* New York: Harper Brothers.

Montias, J.-M., Ben-Ner, A., & Neuberger, E. (1994). *Comparative economics* (Fundamentals of Pure and Applied Economics Vol. 57). Chur, Switzerland: Harwood Academic.

Ofer, G. (2005). Abram Bergson: The life of a comparativist. *Comparative Economic Studies, 47,* 240–258.

Pryor, F. (2008). Culture rules: A note on economic systems and values. *Journal of Comparative Economics, 36,* 510–515.

Resnick, S., & Wolff, R. (1987). *Knowledge and class.* Chicago: University of Chicago Press.

Resnick, S., & Wolff, R. (2002). *Class theory and history: Capitalism and communism in the U.S.S.R.* London: Routledge.

Schnitzer, M. (2000). *Comparative economic systems* (8th ed.). Cincinnati, OH: South-Western College.

Smolinski, L. (1969). Lenin and economic planning. *Studies in Comparative Communism, 2,* 96–114.

Sweezy, P. (1976). *The transition from feudalism to capitalism.* London: New Left Books.

Mises, L. von.(1920). Economic calculation on the socialist commonwealth. In F. A. von Hayek (Ed.), *Collectivist economic planning* (pp. 87–130). London: Routledge & Sons.

Ward, B. (1980). Comparative economic systems: Changing times—changing issues. *American Behavioral Scientist, 23,* 393–414.

Zimbalist, A. (Ed.). (1984). *Comparative economic systems: Present views.* Boston: Kluwer Academic.

Zimbalist, A., Sherman, H., & Brown, S. (1989). *Comparing economic systems: A political-economic approach.* New York: Harcourt Brace Jovanovich.

# 44

# World Development in Historical Perspective

Douglas O. Walker
*Regent University*

Since the end of World War II, the world economy has been transformed from a devastated and fragmented assortment of national economies primarily engaged in producing a low level of basic commodities for local consumption into a unified set of product, factor, and financial markets of global dimensions generating a tremendous volume and variety of goods and services supporting a much higher level of living for the world's population. The combination of high population growth, a brisk pace in capital formation, ever widening markets, and accelerating technological advances led to a rapid rise in production with corresponding large-scale changes in productive structures and patterns of international trade. More than a half century after this transformation began, in the early years of a new millennium, the world economic landscape is markedly different from the scene of destruction, war weariness, and backwardness presented at the end of World War II, with a fully integrated group of prosperous, economically advanced countries tied closely to a group of developing countries rapidly modernizing their economies and, at a distance, mutually bound to a less integrated set of poorer and lagging countries divided by circumstance, culture, polity, and policy.

This chapter reviews the changes that have taken place in the world economy during the second half of the twentieth century into the first decade of the twenty-first. It begins with a brief review of changes in the arrangements supporting world development during this period and the advances that have been made in lifting world incomes and improving levels of living of the world's population. It then focuses on the pace, pattern, and stability of world economic growth from 1950 to the present and notes that it slowed over time and became increasingly marked by disparities and instabilities. The discussion focuses on four characteristics of contemporary world development: the strong recovery in world production from the setbacks of two worldwide wars and the deep depression of the first half of the twentieth century, the rapid increase in international trade in the postwar period despite the widening imbalances and increased instability that now describe the international economy, the rapid initial pace of world growth followed by a marked slowdown in the expansion of world economic activity, and the marked differences in levels of living and growing disparity in individual country growth rates between higher income and lower income areas that describe recent decades. The chapter concludes by noting some challenges before the world economy in a new century.

## Changes in the World Economic and Political Landscape

The dimensions of the economic progress mankind has made since 1950 are reflected in the enormous gains recorded in population, production, and productivity across the entire globe. The Earth can now support a much higher population at a much higher level of living than any previous epoch in human history. Its habitable area has been extended greatly, and huge new areas have been opened to the cultivation of food and the extraction of raw materials. The exploration of the moon, planets, and stars has begun, and in an answer to dreams from time immemorial, man walked on this planet's nearest celestial neighbor. The study of space and the contribution it can make to improving human welfare has begun in earnest and promises great benefits to mankind. The deep oceans now

contribute greatly to meeting humans' everyday needs. Most important, there is not one area of the world that has not gained from the progress humankind has made during this extraordinary period of world history.

These advances were supported by accumulating knowledge about the world, a much more skilled and educated population, and an accelerating pace of technological change shared across borders and encouraged by a strong upsurge in world trade and foreign investment. As the momentum of world commerce increased and outpaced the rise in domestic output and incomes, the production of sophisticated manufactured products has spread from the old industrial centers of Europe and North America to emerging countries where the latest techniques are now being integrated into the economies of previously low-productivity primary-producing countries. Reflecting much higher average incomes, structures of production and patterns of expenditures everywhere are markedly different today than in the past. The geographic distribution of humankind has also changed radically as people have become more mobile nationally and internationally, and human settlements have become increasingly concentrated in high-income urban areas connected by extensive transportation and communication links that tie the world together.

In terms of demographic characteristics, economic circumstances, and social development, the world today reflect tremendous improvements over the situation prevailing in the decades and centuries prior to the mid-twentieth century, especially the troubled period leading up to unprecedented global expansion that has taken place since the end of World War II. This progress reflects not only an increased knowledge about the world but a restructuring of the global economic and political order that promoted international commerce and encouraged better economic management at the national level in the context of a set of generally accepted rules, regulations, and procedures applied to an increasing number of the world's countries.[1]

## Renewal of Economic Advance and the Rebuilding of the International Order

The years following 1950 witnessed a strong renewal of global progress from a protracted and deep worldwide depression and the aftermath of a devastating full-scale war extending across several continents. In the decades following the Great Depression and World War II, despite extraordinary political tensions and the potential for nuclear devastation, many areas of the world entered a prolonged period of peace and prosperity as the global economic and political order was adapted to new realities. Although incomplete, and with many setbacks and glaring disparities among countries, in historical perspective, the progress made the past half century is extraordinary in terms of problems overcome and wealth generated across the world.[2]

In the economic area, real gross world product per capita increased more than threefold from 1950 to 2008 as labor productivity rose rapidly in response to a broadening division of labor, more capital per worker, and innovations in organization, products, techniques, and transport. High and balanced growth in the more economically developed areas was restored to these countries that compares favorably with any previous period in their history. Rapid and sustained growth emerged in many developing countries, spurred by closer international economic integration among all countries. In many regions of the world, a growing convergence of material well-being at an unprecedented high level describes the changes taking place. The rise in output brought with it not only widespread improvements in material levels of living and a better quality of life for people on every continent but also significant improvements to demographic characteristics and great strides in political and social development across the globe. The extent of economic progress was unprecedented in the history of mankind, both in terms of the pace of the advance and in terms of its widespread reach.

In the international area, great progress was made during the past half century. These years saw several periods of major restructuring of the political and economic foundations of all groups of countries that completely transformed the prewar global landscape. Even before the end of World War II, a new international economic order was established—the Bretton Woods system.[3] This system, as adapted from time to time to new circumstances, reintroduced a fixed exchange rate system supported by a set of international economic institutions that served as the basis on which an increasingly globalized economy arose from the ashes of World War II. Indeed, the steady reduction of trade barriers and the strengthened institutional mechanisms for multilateral trade and payments have been a key factor promoting both world economic prosperity and world peace. At the national level, governments increasingly took the lead in creating the conditions for modern economic growth, and institutional structures supportive of growth—legal, educational, and health systems, for example—were introduced where they had not previously existed and strengthened where they had.

The prewar decades also saw the creation of two competing national economic systems, one based on the market mechanism and one based on centralized planning, and the emergence of deep political divisions between East and West and North and South. But differences between East and West ameliorated with the passage of time, and it would seem clear that the sweeping economic and political changes that have taken place in central and eastern Europe and the states of the former Soviet Union in the 1990s have the potential of moving the world closer to a stable political order and to a worldwide economic system based on a single set of commercial and market-oriented principles. Similarly, differences between North and South lessened as trade expanded and incomes rose in many countries of the South. In this regard, an ongoing

reassessment of development policy orientations brought many developing countries to the same view as the economically advanced countries: that their objectives are best achieved by greater reliance on markets and the international economy. This debate about the best strategy to accelerate development and manage economic prosperity is still under way, however, and will take many decades to resolve.

With regard to the developing countries, the decades of the 1950s and 1960s saw another set of changes as the global political order was rearranged to accommodate the emergence of the former and dependent territories of the European powers as independent nation-states. The adjustment on the part of the European countries was so effective that by 1970, no large dependency state remained under their control, yet European influence, culture, and commercial ties remained strong. Adjustment on the part of some developing countries, left with the problems of governing societies undergoing fundamental change, has been less successful. Failure is reflected in the setting up of weak and ineffective governments in many countries and continuing strife in southeast Europe and central Africa and in central, south, and western Asia, as well as numerous festering contests over the economic and political order in other parts of the world. Political fragmentation and economic instability represent a legacy the twentieth century leaves to the twenty-first.

Failure is also reflected in the hostility in some quarters against the multilateral institutions promoting international economic integration and multinational corporations carrying the latest technology and newest products to the most remote corners of the globe. Much of this hostility reflects difficulties in economically advanced countries when adapting to changes brought about by technology and trade and pressures on developing countries to give greater emphasis to the importance of markets and the gains from international exchange. Indeed, this hostility goes far beyond economic issues of trade, technology, and adjustment. At its root, it reflects the fear that traditional livelihoods and cultural institutions in every country are threatened by the process of globalization and the changes it brings in its wake. At a deeper level, it reflects opposing views about the organization of society, with a philosophy of spontaneous, market-led coordination of mankind's activities competing with one of collectivism and statism where government, community, or religion orchestrates production and consumption and sets the agenda of daily life.[4]

Nevertheless, while it is true that a debate about the goals of development and the best strategy to attain them is under way, one should not underestimate the capacity of cultures to absorb external influences and their ability to adapt their most deeply held religious and cultural traditions to new circumstances and innovative ideas. In fact, tremendous changes have already taken place in the cultural and religious sphere of all countries in response to the process of modernization and globalization. Many of these changes are unwelcome and therefore remain subject to revision and reversal as each society's norms and aspirations absorb and adapt superior influences from abroad and modify and reject what they regard as inferior currents. Nevertheless, institutions, both formal and informal, conducive to economic advance have emerged everywhere, and their influence is likely to increase in the decades ahead.

In retrospect, the enormous changes in the economic, political, and social realms that have taken place in the past 60 years, any one of which could well have destabilized the global political order and set back world economic progress for decades if not centuries, were introduced with minimal disruption. It must also be noted that however horrible present conflicts are to those engulfed by them, and however deep the poverty and however great the contention over prevailing cultural, political, and religious norms, they are not comparable to the squalor and deprivations of the past or to the horrors of the worldwide slaughters of the first half of the twentieth century. It is difficult not to conclude when looking back at the past half century, keeping in mind the worldwide wars, depressions, and disintegrating political arrangements of the first half of the twentieth century, that the political development of the world and its peaceful transition to modern economic arrangements at the national and international levels was remarkably good after 1950, especially considering the challenges the world faced and the history it had to build upon.[5]

## Accomplishments Since 1950

In the area of social and economic accomplishments since 1950, the list is long. The real gross world product, measuring the aggregate volume of economic activity produced by all people in all countries, has risen by a factor of more than six and one half since that year.[6] This long-term rise in the capacity of the world economy to supply goods and services has been accompanied in all main world regions by significant changes in patterns of resource use and structures of production, in population dynamics and labor force characteristics, and in social conditions and possibilities for human interaction. As a result, living standards have risen greatly over the past 60 years, and absolute poverty has been significantly reduced globally and eliminated in many countries. International trade with its deepening commercial integration has risen at an even faster rate than output, and we now live in a world where production is global, foreign travel is commonplace, and factor mobility across countries is accelerating.

The rise in output has brought with it not only widespread improvements in material standards of living and a better quality of life for much of the world's population but also great strides in social development. A few of the more important changes in this area may be listed as follows:

1. The increase in world output has supported not only a great rise in world population and labor force but a great rise in output per person (see Figure 44.1).

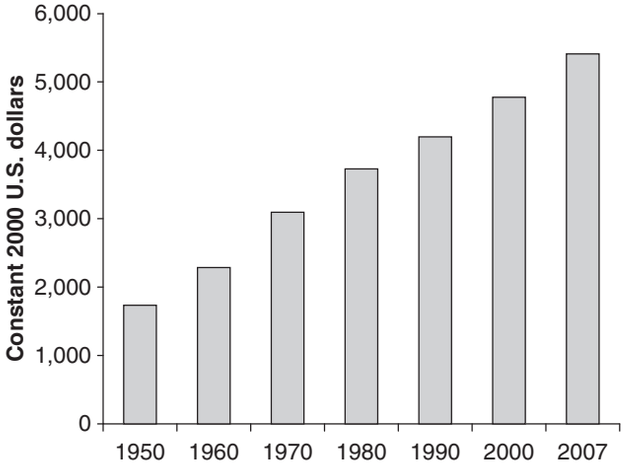


**Figure 44.1** Rise in World Per-Capita Product

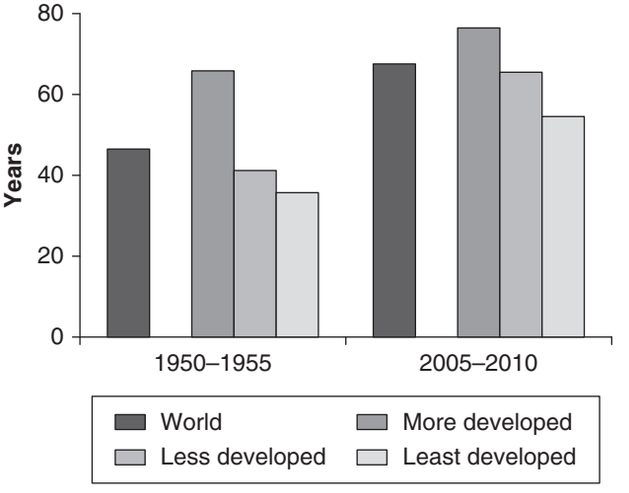


**Figure 44.2** Rise in World Life Expectancy

2. Marked reductions in mortality have been recorded, which raised average world life expectancy by two decades (see Figure 44.2).

3. The reduction in mortality was accompanied, but not simultaneously, by a decline in fertility, with a fall in crude birthrates from a figure generally over 40 per 1,000 to well below 15 in some countries (see Figure 44.3).

4. School enrollment ratios increased from low and sometime negligible figures to ratios that included most if not all children, with a similar extension of education to previously ignored groups of youth and adults. The improvement in literacy is particularly notable in developing countries, where education for boys is nearly universal. The situation with regard to girls is less good and requires more attention (see Figure 44.4).

5. Large changes in labor force characteristics can also be noted in response to changes in the age distribution of the population, in propensities of women to enter and exit the labor force, and, most important, to investments in human capital (see Figure 44.5).

Remarkable improvements in health status have taken place everywhere in response to new medicines, improved diagnoses, and greater access to health services. In turn, a better educated and more mobile population enjoying a longer and healthier life has provided the basis for continued gains in productivity permitted by a better fed, better educated, healthier, and more migratory labor force.

And it should not be forgotten that hand in hand with improvements in direct measures of well-being have come continuing advances in technology—especially in telecommunications and travel, in the biological and material sciences, and in information and computer technology. These improvements not only have intensified economic interdependence among countries but have also had a profound effect in each nation's cultural sphere as people have become increasingly aware of events, lifestyles, and ideas in the rest of the world. Continuing advances in technology promise further increases in productivity and standards of living to the entire world's population.

Given this experience, there can be no doubt but that the present era is one of great human advance that has been sustained—despite a number of persistent difficulties and setbacks of high cost—for more than a half century.

## World Economic Performance in the Postwar Era

Among the most important factors affecting the life of each person on the planet is the performance of the world economy as it is reflected in the pace, character, and distribution of economic growth among its many countries. A high rate of economic growth is seen as a primary determinant of a country's ability to raise the level of living of its inhabitants; provide them with more meaningful kinds of employment, better educational opportunities, and improved health status; and maintain and enhance the physical environment in which they live. At the same time, rapid economic growth necessarily involves a large-scale mobilization of resources and extensive structural change, processes that often involve social disruptions and unsustainable elements such as ecological damage, the accumulation of wastes and pollutants, and the profligate use of energy, forests, water, and other natural resources. Moreover, economic advance always challenges existing political and social arrangements, often with the loss of cherished traditions


**Figure 44.3** Changing World Demographics

and ways of life that people value more than higher incomes and wider economic opportunities.

For this reason, the benefits world growth brings must always be matched against its noneconomic costs. Nonetheless, it remains true that without a steady and continuing rise in income per person, the capacity to sustain, much less improve, living conditions for the vast majority of the world's population who continue to live in dire want is impossible. Therefore, any assessment of the long-term performance of the world economy must also include the contribution growth has made to raising levels of living across the entire globe, especially in the world's poorest countries where abject poverty remains endemic and an unacceptable blight on the human condition.[7]

## Six Decades of Rising World Output

Without question, an unprecedented period of global economic expansion and diversification began after World War II. Output in all world regions has risen markedly and for the most part continuously since 1950, and the rate of increase recorded—an average rise of about 3½% each year—compares favorably with any past era in human history (see Table 44.1). World population rose in numbers from 2½ billion in 1950 to more than 6½ billion today, an

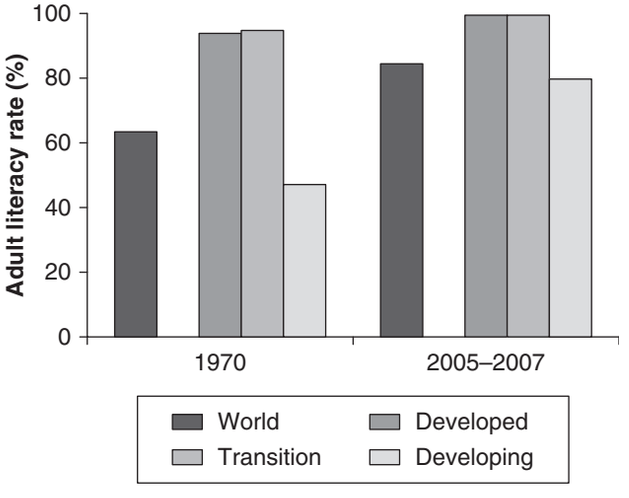


**Figure 44.4** Rise in World Literacy

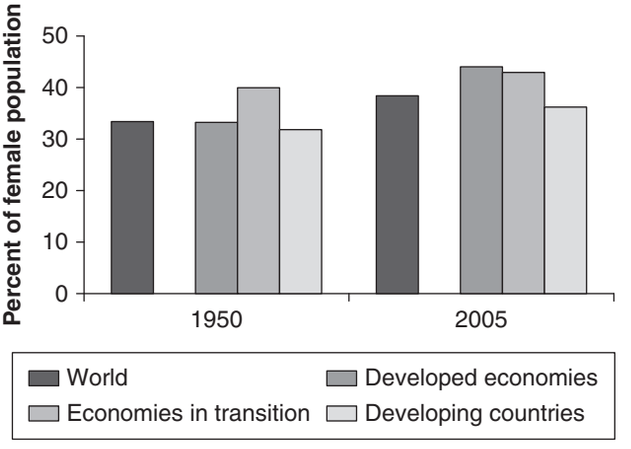


**Figure 44.5** Rise in World Female Labor Force Participation

**Table 44.1** Growth of World Output, Population, and Labor Force, 1951–2007 (Average Annual Rates of Increase)

| | *Gross Product* | *Population* | *Labor Force* |
|---|---|---|---|
| World | 3.4 | 1.7 | 1.8 |
| Developed economies | 3.0 | 0.8 | 1.0 |
| Economies in transition | 2.6 | 0.6 | 0.6 |
| Developing countries | 5.0 | 2.0 | 2.2 |

annual increase of 1¾%, about double the rate of increase from 1900 to 1950 and triple the rate of the century before. The world labor force rose even faster than population as changing social norms and greater control over fertility increased women's participation in the workforce.[8]

With the more rapid increase in production over population have come significant improvements in levels of living for most of the world's population as well as an increased capacity to address environmental, health, and social problems. Spurred on by rapid growth, rising international trade, and rapid advances in technology, the structure of the world economy has also been transformed radically as essentially rural and traditional economies became urban and more commercially oriented producers of goods and services for domestic and foreign markets. With this transformation have come better employment opportunities, safer places of work, and less time spent in the workplace.

However, the rise in world output and its distribution among the world's nations has not been uniform over time or over the globe. In the industrialized and service-oriented economies of Europe, North America, Japan, and Oceania, which taken together account for the preponderant proportion of organized global economic activity, this long-term expansion has been characterized by widespread prosperity and a process of convergence in which the initial gap in productivity among these countries has been progressively reduced over time. Although these countries enjoyed a long-term and broad-based expansion, growth in the more economically advanced countries was nonetheless interrupted several times by deep recessions, and the functioning of their economies over time reflected increasingly high unemployment, growing imbalances, and difficult recoveries. More significant for their future, as discussed below, the average pace of economic growth in many of these countries tended to slow down noticeably over time, and widening disparities in income distributions have come to characterize many of them in recent decades (see Table 44.2).

At the end of World War II, the countries of Eastern Europe and the former Soviet Union faced a heavy burden of reconstruction and undertook far-reaching changes in the organization of their production and trade. Seen in longer term perspective, the tempo of production in East Europe and the successor states of the former USSR, which had been rapid for three decades under central planning, slowed markedly during the 1980s, and the absolute level of real output produced in these countries fell precipitously in the 1990s, when the economic model on which their production had been based proved unable to sustain their level of living. These countries are now undergoing a remarkable political and economic transition from state socialist and centrally planned economies to a more market-oriented and open economic system with greater integration into the world economy. Growth in many transition countries in the 2000s has been bolstered by high commodity prices and is now higher than rates recorded in earlier decades. Nevertheless, their share in world production is lower today that it was a half-century ago, and the distribution of income within these countries has become more skewed.

The past six decades have also witnessed great progress in the economically developing countries of Africa, Asia, and Latin America and the Caribbean, but this advance has been combined with setbacks in some countries and regions: most notably, lagging growth in many of the world's poorest countries. Generally rapid and rising economic growth over the entire period has been recorded in South and Southeast Asia. This growth has

**Table 44.2** Share of World Output, Population, and Labor Force, 1950 and 2007 (Percentage Share in Corresponding World Total)

| | *Gross World Product* | | *World Population* | | *World Labor Force* | |
|---|---|---|---|---|---|---|
| | *1950* | *2007* | *1950* | *2007* | *1950* | *2007* |
| World | 100 | 100 | 100 | 100 | 100 | 100 |
| Developed economies | 83.7 | 70.7 | 22.7 | 13.5 | 23.3 | 14.3 |
| Economies in transition | 3.7 | 3.5 | 10.6 | 6.1 | 11.0 | 6.2 |
| Developing countries | 12.6 | 25.8 | 66.7 | 80.4 | 65.7 | 79.4 |

been linked to international specialization and the advantages offered by cooperation within the established frameworks for open multilateral trade. In China, an extraordinary growth momentum has developed over the past three decades, with sustained increases in production on the order of 8% to 10% a year. In Latin America and the Caribbean, on the other hand, early decades of expansion have been replaced by decades of stagnation and slower growth performance as this region responded with great difficulty to external shocks that arose during the 1970s and 1980s. Similarly, in Africa and Western Asia, an initial period of good performance has been followed by an extended period of weak and faltering growth, which has recently strengthened in the global economic upturn of the early years of this century but remains fragile and dependent on continuing high commodity prices.

In the opening years of the new century, a more rapid pace of economic growth returned to all developing country regions. To sustain this growth, many of these countries must address the legacies of persistent internal and external imbalances and high foreign debts inherited from the past quarter century. The difficulties of this adjustment are compounded by the challenge presented by the very process of development and the changes that must be introduced to adjust and adapt to pressures of globalization and the tendency toward growing disparities within their national economies. This adjustment has been made more difficult by the 2008–2009 downturn in the global economy but is essential for creating conditions for long-term world growth.

## A Buoyant but at Times Difficult International Economy

The global economic expansion of the past six decades benefited from strong international institutions and a buoyant international economy, generally supportive national economic policies that recognized the importance of the external sector, and an unprecedented pace of technological advance that encouraged a wider division of labor among the world's economies.

The unstable international economy of the interwar years led to the recognition that the foreign trade sector would be critical for all countries in the postwar era. For the first time, the Bretton Woods agreement of 1944 established a formal international system to govern monetary and trading relations among independent nation-states. When designing the system, the objective was to promote an open multilateral system of international trade based on fixed exchange values, a steady flow of international liquidity, mechanisms to support the balance-of-payments adjustment process, and rules governing the conduct of international trade. To this end, the Bretton Woods system called for the gradual removal of trade controls and impediments, as well as the attainment of sustained rates of domestic economic growth by raising the productivity of the national workforce of each country through participation in a wider international division of labor.[9]

The international arrangements and more outward-oriented policies of the postwar era brought about an expanding flow of world trade that stands in marked contrast to the sharp contraction recorded during the prewar years. An extraordinary expansion of international commodity trade and a great increase in cross-border travel, communications, and exchange of other services began after 1950 and have continued to the present. There were, of course, setbacks, and not all countries and regions participated in the growing international economy. But except for a turbulent period in the mid-1970s and the early 1980s, the reverses were few and of short duration, and while the momentum of trade lessened somewhat after several decades, the strong upward trend in international trade has been a major factor promoting and sustaining world economic growth.

In terms of the tempo of the expansion, the increase in world trade has been remarkable. On average, world trade in goods and services has risen at a rate of 6% each year since 1950, much more rapidly than the increase in world output and much faster than in any previous historical period of similar length (see Figure 44.6). In the decades following 1950, the volume of world trade, measured in constant dollar prices, increased by a factor of almost 35, whereas the volume of world output rose by a factor of only 8½. With the much faster rise in world trade over domestic product has come a much more closely integrated world economy, with tighter and wider interdependencies among its many countries and a growing awareness that the world is becoming globalized not only in production and exchange but in culture and social development.

Although widespread, the growth of the international economy has not been even over the world. Regions of the world that have participated actively in international commerce, such as the more economically advanced countries of Europe, North America, Japan, and Oceania and many developing countries in Asia, have experienced extraordinary increases in their trade over a long period of time. Countries in other world regions that chose not to fully participate in the international economy have found that the increase in their trade has been halting at times and characterized by turbulence and great instability. In the case of developing countries, the pace of trade expansion has been on the order of one and one half to two times as fast in Asia than in other areas of the world. Given these disparate rates of expansion, the developing countries as a whole, which accounted for approximately one quarter of the value of world trade in 1960, increased their share to more than one third a half-century later (see Figure 44.7).

Although rapid, the growth of the international economy has not been steady over time. Indeed, despite its overall vigor, at times it has been halting and characterized by turbulence and great instability. At the international level, following several decades of strong increases in world trade, a weakening of the international monetary and trading system occurred in the late 1960s and 1970s, reflected in volatile and misaligned exchange rates, fluctuating commodity prices, difficulties in balance-of-payments adjustment, and

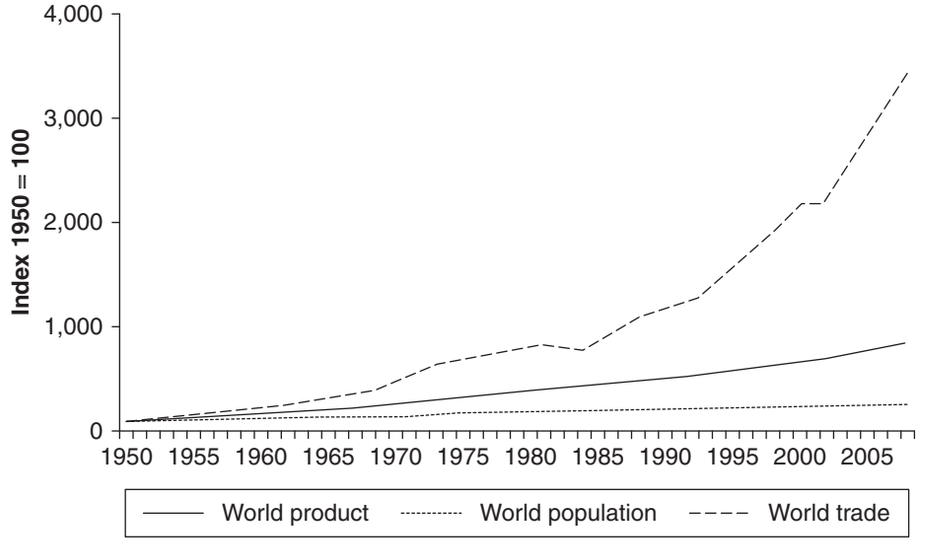


**Figure 44.6** Rise in World Output and Trade

accumulating external and internal imbalances, especially growing external debts of developing countries. These problems have periodically caused interruptions and instabilities in the upward trend in world trade that have continued to the present (see Figure 44.8).

The pressures on the original Bretton Woods system during the late 1960s and early 1970s were particularly great. Toward the end of the 1960s, accelerating inflation, a widening external deficit, and an expanding pool of dollars accumulating as reserves and in banks outside the country caused the United States to terminate the convertibility of the dollar into gold in 1971. The upswing in world production and trade from mid-1971 into 1973 was one of the strongest on record, but it was associated with increasing instability in international commodity markets and growing pressures on domestic productive capacity. Inflation in the international economy accelerated in response to strong demand pressures from rapid world growth at a time of supply losses due to poor harvests. In 1973, the stability of the world economy began to erode further when the U.S. dollar came under speculative pressures, and the worldwide system of fixed exchange rates established at Bretton Woods came to an end when a floating currency regime was introduced. Adding to the problems, later that year, in the unsettled

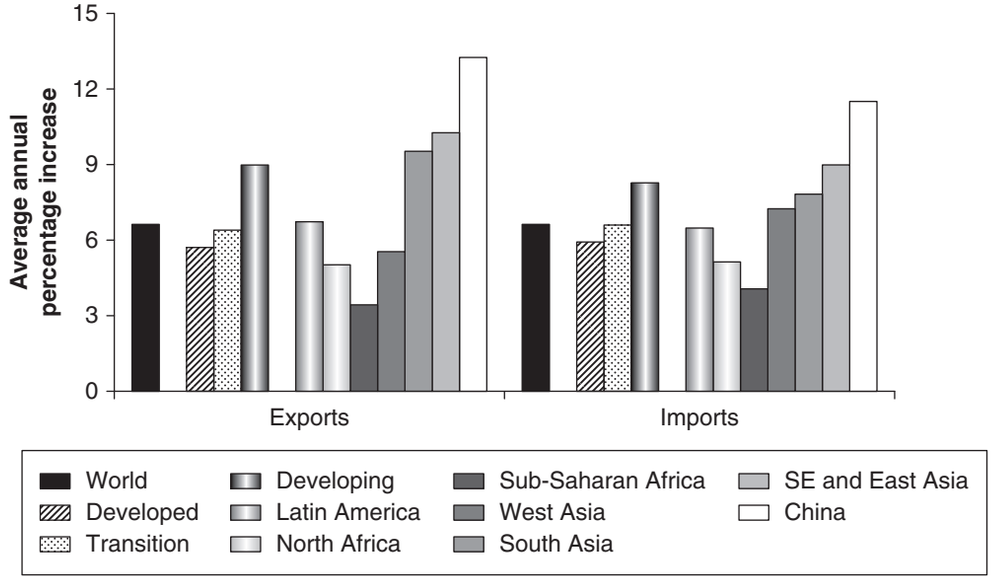


**Figure 44.7** Growth of World Trade

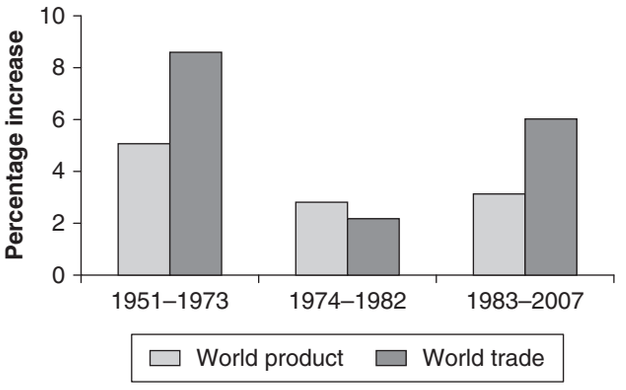


**Figure 44.8** Growth of World Output and Trade by Period

environment, the dollar price index for commodities entering world trade rose greatly, with the Organization of Petroleum Exporting Countries increasing crude petroleum prices fourfold within a year. As a consequence, the environment for trade and world growth became much more adverse, and twice a sharp decline occurred in the real volume of international trade, once in the mid-1970s and again in the early 1980s.

In retrospect, a remarkable era of world economic growth came to an end in 1973 as the tempo of global production and trade thereafter slowed noticeably and exhibited wider fluctuations and greater disparities in national economic performance. For the first time in the postwar period, the trend rate of world trade growth fell below the pace of world output as the world experienced its first major downturn since the years of global recovery and reconstruction directly after the war. In the wake of the collapse of the Bretton Woods system, wide deficits and surpluses came to describe the balance of payments of different groups of countries. The instabilities of the first half of the 1970s not only ushered in the first major recession of the postwar period but fundamentally changed the international price relationships and set of currency relationships on which the previous pattern of world growth and trade had been based. Throughout the 1970s and into the 1980s, higher commodity prices not only fueled inflationary pressures but also changed the basis for industrial costs that required a significant adjustment to domestic patterns of economic activity and resource use. Similarly, changes in exchange rates magnified the need for new and significantly different economic relationships among countries as their former focus of international specialization became obsolete in a markedly different international economic environment. While long-term world growth recovered somewhat after the early 1980s, it has not reattained the momentum recorded from the 1950s to the early 1970s.

All major groups of countries and all geographic areas have benefited from a dynamic international economy of the past six decades. It has at once boosted their possibilities for growth through greater specialization, leading to higher productivity and a wider variety of goods and services, leading to higher levels of living. But at the same time, increased participation in the world economy has brought economic and financial difficulties of sizable proportions to each and every country. With greater international exchange have come intensified pressures for adjustment and adaptation to rapidly changing patterns of international commerce, and increasingly, integrated world product, factor, and financial markets have increased each country's vulnerability to events and conditions occurring in the rest of the world. While difficult and presenting major challenges to all countries, there can be no doubt that the world has benefited greatly from the more stable and prosperous international economy of the past half-century than the hostile and protectionist international economic environment that described previous epochs in world economic history.

## Acceleration Followed by Slowdown in World Productivity Growth

The efficiency with which the world uses the capital, labor, and natural resources at its disposal is the most important determinant of the level of living of its inhabitants and their possibilities for addressing the economic problems humankind faces. For this reason, the increase in the productivity of each person in the workplace is a key consideration when assessing how successful the production process has been in improving the lives of people because the growth of consumption per capita is related to the increase in output per worker. Because the number of people in the labor force tends to move with the number of people in the population, labor productivity growth measured as output per worker drives the rise in real incomes per person over time.[10]

Seen in long-term perspective, the rate of world productivity increase has been exceptionally high during the last half of the twentieth century into the twenty-first. World output per person in the labor force has been estimated by Angus Maddison (2001) to have risen at most 1% a year on average in the decades immediately prior to 1950 and an even slower pace in earlier centuries. In contrast, the pace of world productivity advance accelerated to an average annual increase of 1% a year in the decades after 1950, a historically unprecedented rate of growth for such a long period and one that would double average levels of income in about 40 years (see Figure 44.9).

Reflecting this fast pace of change, world output per economically active person, measured in 1990 prices and exchange rates, rose from an average of $3,760 in 1950 to more than $11,400 in 2007. International trade and rapid advances in technology promoted this rise in overall productivity through a wider division of labor at the world level and a wider range of goods and services in the domestic marketplace. As a consequence, as world levels of living rose, domestic economies were completely transformed in terms of their productive structures as increases

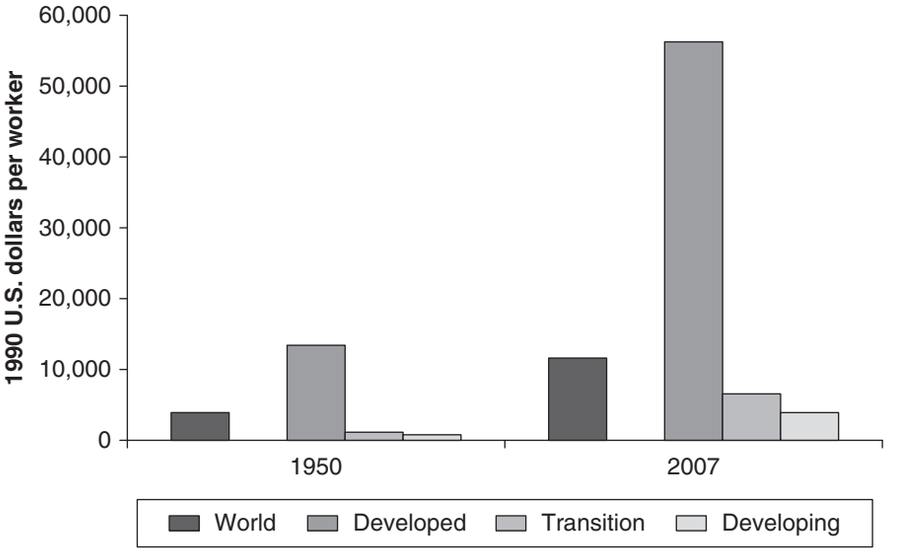


**Figure 44.9** Rise in World Labor Productivity

in output per person in agriculture and manufacturing not only far exceeded those in services but rose faster than the growth of demand for these products. What was mainly an agrarian and industrial world economy in 1950 became a world economy increasingly oriented toward the delivery of services.[11]

However, it is also fair to say that, although unprecedented, in some respects, long-term world productivity performance has been characterized by deficiencies that became increasingly evident over time (see Figure 44.10). The broadening increases in productivity that did emerge in the early postwar decades, for example, should have been sustained and strengthened in later decades as a gradual diffusion of technology to all countries and a wider scope for introduction of new methods of production in low-income countries supported strong, steady, and spreading rates of economic growth. Lower transport costs, falling trade barriers, and what would appear to be cascading economies of scale spreading through the world production system should have supported not only a widening division of world production and a far greater variety of goods and services available at the national level but also an accelerating expansion of output as average costs declined continually and demand climbed steadily in response to higher real incomes. In the same manner, increased attention to investment in human capital in all countries, especially in low-income countries, should lead to the entrance of a growing number of workers with enhanced educational backgrounds and training, thereby raising the knowledge and skills of the labor force. Finally, rising life expectancy and an improvement in a wide range of social conditions in almost all countries should also provide a basis for a rapid, equitable, and steady expansion in world economic activity.

Contrary to expectations, however, accelerating technological advance, a widening network of international trade and investment, and rising literacy and improving social conditions in all groups of countries did not raise—indeed,


**Figure 44.10** Long-Term Decline in World Productivity Growth

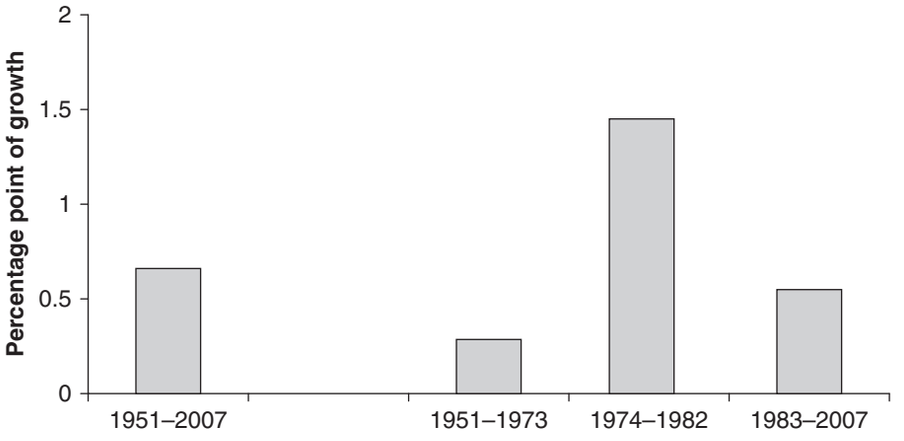


**Figure 44.11** Swings in Annual World Productivity Growth

did not even prevent a persistent fall in—the long-term rate of world economic growth.

While the early period of the 1950s, 1960s, and into the 1970s were years of high, stable, and balanced worldwide growth, beginning in the early 1970s, the pace of world productivity advance showed a persistent tendency to slow down and exhibit greater instability and disparities. For much of the earlier period, the international monetary system provided the needed international liquidity, and the international adjustment mechanism allowed for the effective elimination of balance-of-payments disequilibria without resorting to overly restrictive domestic policies and the introduction of barriers to trade at the national level. Investment effort—the share of produced resources devoted to physical capital formation—rose, and investment efficiency—the lowest number of dollars of investment necessary to raise labor productivity by a dollar—remained high. As a result, from the early 1950s into the early 1970s, the rate of world productivity growth was rapid, increasing at an average of more than 3.3% a year, truly an outstanding rate of increase in productivity for such a long period of time (see Figure 44.11).

However, the high and steady expansion of the early postwar years was followed by slower and more unstable growth in later decades. As noted earlier, at the national level, widening imbalances in the external accounts came to be recorded in many countries. Inflationary pressures also increased over time, with the rise in U.S. dollar–weighted prices accelerating steadily from less than 3% a year in the 1960s to more than 14% a year by the end of the 1970s. Marked accelerations and decelerations in economic growth were recorded as short-term trends in industrial production, and wide swings in aggregate demand led to steep recessions in the mid-1970s and early 1980s. In the unstable economic environment of these years, rising unemployment and slower trend world productivity growth could also be noticed as the decade of the 1970s ended and that of the 1980s began (see Figure 44.12).

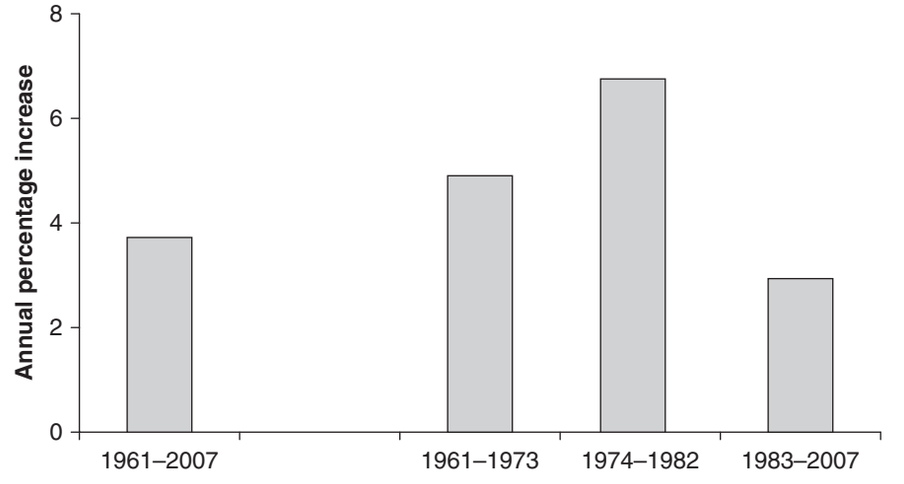


**Figure 44.12** Long-Term Trend in World Inflation

Although world productivity growth revived somewhat after the early 1980s, and progress was made in reducing inflation from the double-digit rates of the late 1970s, the pace of world productivity growth the past quarter century is half the rate recorded the first quarter century after World War II. Similarly, unemployment rates are significantly higher than in the past, and while the rate of world inflation has fallen to a point today where deflation has emerged as a concern, seen in longer term perspective, U.S. dollar prices are much more volatile than in the past. Wide imbalances have arisen in the external accounts of many countries, and exchange rate fluctuations remain a concern for many countries. Even given these trends, it cannot be denied that the past 25 years have been remarkable for the world economy.

## Growing Disparities in Productivity and Levels of Living

Given the rapid, if slowing, pace of world growth over the longer term, there are reasons to believe that the increase in world output and overall improvement in living standards should be shared in an equitable manner among richer and poor countries. In the standard paradigm of long-term economic growth, the expected pattern of world development is a path of convergence in productivity among countries as the rate of increase of countries at lower levels of productivity exceeds the rate of increase of countries at higher levels of productivity. This tendency toward convergence derives from an assumption of diminishing returns to capital, where levels of output produced per person in countries with more capital per worker tend to have lower rates of return on investment and lesser increments in output produced per worker than countries with less capital and lower levels of productivity. In this view, as capital accumulation takes place in high-income countries, the productivity of capital falls, and the incentive to save is reduced; conversely, the higher productivity of capital in low-income countries encourages saving and capital accumulation, thereby generating a more rapid pace of economic growth from any given share of investment in gross domestic product (GDP). As a result, poorer countries, with less capital per worker and higher rates of saving, should grow faster than richer countries as they reap higher returns from investment in plant, equipment, and other physical assets.

For this reason, it was expected that the wide and growing disparities in income and wealth among different regions of the world that existed in 1950 should have progressively lessened as economic growth spread across the globe. It is possible that some widening of income gaps between rich and poor could occur if some countries take off into modernity before others. However, as more and more countries begin the process of modern economic development, these gaps should tend to close as countries at low levels of income per capita grow faster than countries at high levels of per capita income. Moreover, it was

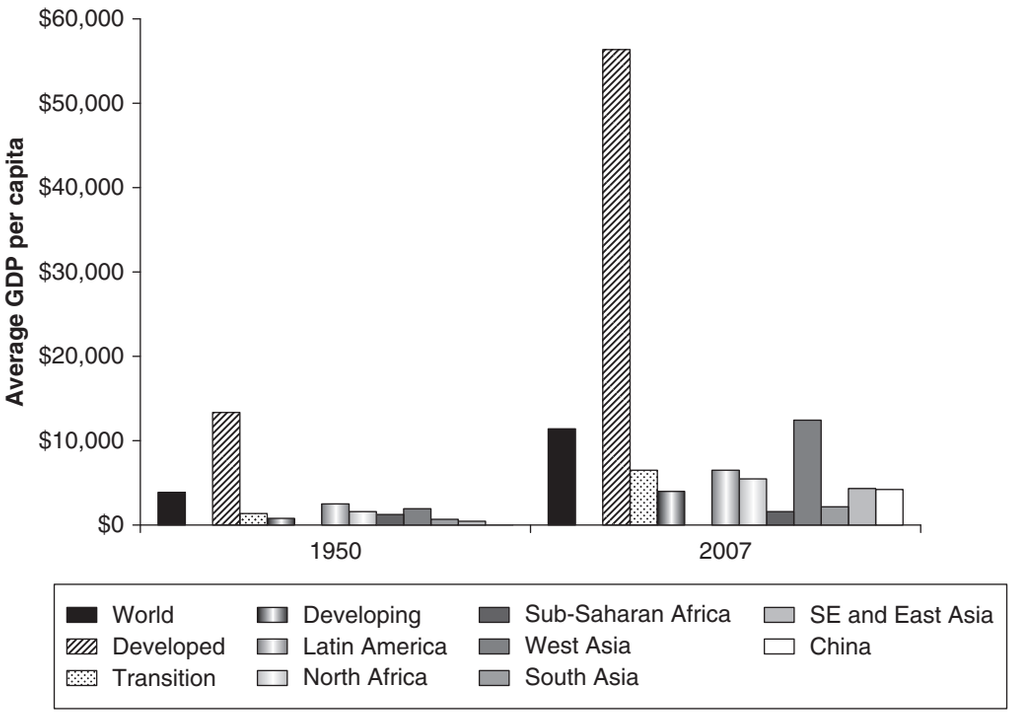


**Figure 44.13** Rise in World Productivity, 1950 and 2007

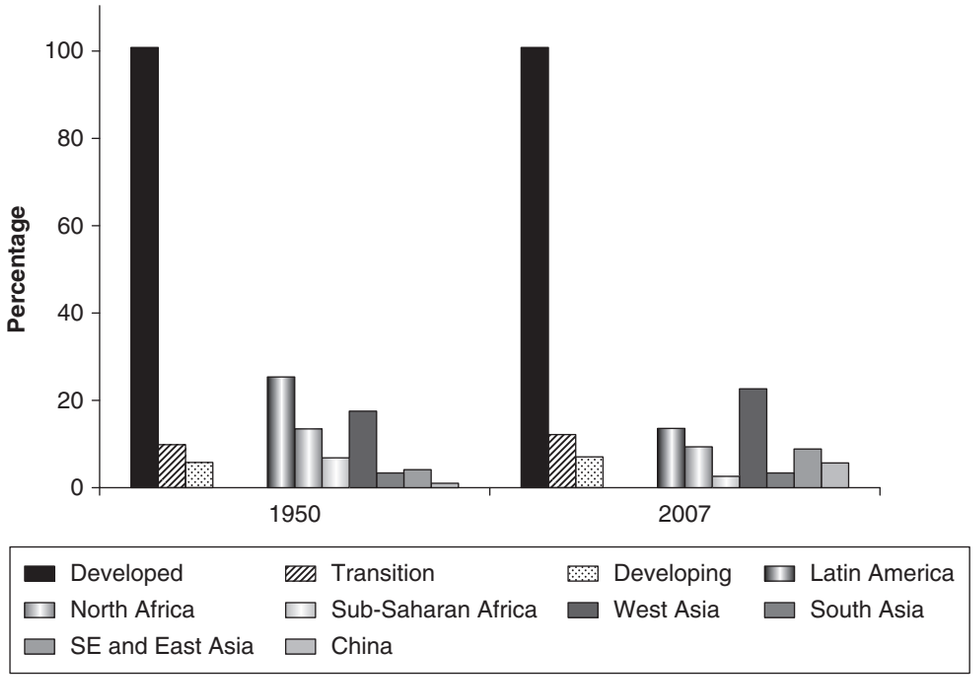


**Figure 44.14** Disparities in World Productivity, 1950 and 2007

thought that countries at lower levels of technological development could draw on the advances already in place elsewhere to accelerate their growth. Over a period as long as half a century, a clear process of convergence in average levels of income among the world's countries should have begun.

In terms of productivity, when classified broadly by level of development, output produced per worker in 1950 in the more economically advanced areas was far greater than in other world regions (see Figure 44.13). Measured in 1990 prices and exchange rates, it is estimated that average GDP per economically active person was $13,525 in 1950, or more than 18 times higher than the level of output per worker produced on average in developing countries ($720) and more than 10 times higher than that produced in the countries now classified as economies in transition ($1,260). Within these major economic regions, productivity levels differed among the countries and subregions, especially in the case of the developing countries, where the range in average output per worker in 1950 was as high as $3,400 in Latin America and the Caribbean and as low as $150 in China.

More than half a century later, productivity levels had risen greatly in all major groups of countries, and some narrowing of the income gap among the broad groups of countries can be seen. By 2007, productivity in the developed economies had risen to an average of more than $56,000 per worker, or a factor of more than 4. Taken as a group, the relative rise from the levels of 1950 was even greater in the case of the developing countries and the economies in transition, as the average level of output produced per worker in these two major economic regions rose by a factor of 5 (see Figure 44.14). In the case of all developing countries taken together, output per worker had increased to $3,700, and the productivity gap with respect to the developed economies had been reduced from more than 18 to 1 in 1950 to 15 to 1 in 2007; in the case of the economies in transition, output per worker rose to $6,430, and the reduction in the productivity gap was from more than 10 to 1 down to less than 9 to 1. Seen in terms of broad changes across the world, the expansion of the past half-century set in motion a slow process of income convergence across the globe, but great differences in average levels of incomes and productivity between rich and poor countries remained.

Moreover, while the pattern of world growth has closed somewhat the gap between broad groups of countries, growth has not been evenly spread within the group of developing countries. Consequently, the dramatic differences in levels of average productivity and incomes that existed in 1950 have become wider over time. In general, the populous developing countries in Asia have begun to close the income gap, while those in relatively less populous regions of Latin American and the Caribbean and in Africa have lagged in the process of world development. Indeed, over the course of the past half-century, disparities in levels of living between the least developed countries of sub-Saharan Africa and all other countries—developed, in transition, and developing—have become increasingly stark. For these countries, the world expansion of the past half-century has not operated in a way that counters the wide differences in levels of living that existed in 1950.

## Record of World Economic Performance and Prospects for the Future

The record of world economic performance in the postwar era is the greatest in human history. For practical purposes, during this period, all of humankind was mobilized for the cause of economic advance and social betterment. Those parts of the world that had in the previous half-century suffered from a Great Depression and had mobilized twice for a total war recovered to enjoy a high and rising standard of living and a prolonged period of relative peace. In those parts of the world where economic advance before 1950 was limited or where it had never taken root, marked progress was made in lifting incomes and improving the lives of literally billions of people who were previously engaged in a daily life-and-death struggle for survival. These years saw world food production per capita rise steadily, industrialization spread across the globe, commercial services come to dominate production, and the services of government become essential to society's welfare, a complex web of intergovernmental agencies watch over the international economy, and international trade, travel, and communications link the world together. The world economy today is entirely different from that of the past.

Of equal significance to the successes of world development are the legacies of failure and the demands of continued progress. The twentieth century leaves to the twenty-first its failure to accelerate the development of the world's poorest countries so as to ensure balanced world development across all countries and all people. Even where it has taken root, the gift of development has brought with it never-ending adjustment costs attendant to advancing technology and the spread of innovations and new ideas across the world. As Joseph Schumpeter (1950) noted, the costs of what he called "creative destruction"—those involved in introducing new products and structures of production, removing obsolete and unproductive activities, and redeploying their associated resources to sectors of rising productivity—are heavy and increasingly unwelcome. The costs of economic advance are short term, sharp, and concentrated, while the gains are long term, gradual, and dispersed. Understandably, with the comforts of development has come hostility to continued advance on the part of those who now enjoy living standards never before achieved by humankind.

But the vast mass of humanity has yet to benefit from the fruits of the unparalleled economic advance of the past half-century. Despite its costs in terms of the disruption to present patterns of production and trade, a continued and increasingly wider rise in global output is essential to lifting the living standards of the world's poor. This is not simply a moral imperative. If world economic growth is not sustained in a way that spreads economic progress across the planet, conflict among the world's people is inevitable. Equally, challenges of aging populations in rich countries amid continued rapid population increase and high fertility in many poor countries represent major concerns about the future that can be overcome only with the additional resources generated by widespread economic growth. Finally, as the process of world development proceeds, humankind must find a better way to manage the global commons and deal with the most dangerous trends of climate change, species extinction, and unsustainable land and resource use. Making world growth compatible with global environmental sustainability is the greatest challenge of the twenty-first century.

## Notes

1. Easterlin (1998) provides an overview of the process of modern economic growth and the transformation it has brought about since its onset in the mid-eighteenth century.

2. The difficulties before the world economy and the precarious nature of the world economic situation before and immediately after World War II are discussed in depth in Part 2 of Yeager (1966). Quantitative comparisons of world economic progress during this half-century with previous epochs of world economic history are given in Maddison (2001). On the need for the United States to assume a leadership role in the international economy during the Great Depression, see Kindleberger (1973).

3. The conference at Bretton Woods created three international institutions: an International Monetary Fund to supplement inadequate reserves of gold with credit facilities, stabilize exchange rates, and promote more orderly balance-of-payments adjustment; a World Bank to provide long-term capital for postwar reconstruction and international investment in poorer countries; and a stillborn International Trade Organization (many of whose functions were later assumed by the General Agreement on Trade and Tariffs and later strengthened under the World Trade Organization) to set rules governing commercial exchange and liberalize trade. Designed with the problems of the 1930s in mind, these institutions had to be adapted to the emerging problems of the postwar period—most significantly, the growing economic and political importance and problems of developing countries.

4. The debate between classical liberalism—to use the old term for market-oriented economies and democratic-oriented (in the sense of Hume and Locke) political arrangements—and democratic socialism—to use the old term for government-guided economies and democratic-oriented (in the sense of Laski and Rawls) political arrangements—goes back before the turn of the twentieth century. The fortunes of both sides in this debate have waxed and waned for several hundred years in Europe and North America. What is new in the postwar period is that this debate has been extended to the developing countries and, indeed, to international arrangements as well as domestic. The debate also encompasses questions not considered heretofore, such as environmental impacts, equity considerations in factor remuneration, and measures directed at poverty reduction. Issues in this debate are discussed in Henderson (2001). Beyond these issues are much more fundamental questions about societal order that emphasize mankind's relationship to its Creator and the implications that this has for economic and political arrangements. These latter questions are not simply the most difficult to address: They

are also the most fraught with disagreement and the potential for conflict, as events since the September 11, 2001, attacks on the United States show.

5. The historical origins of the global economy from the Industrial Revolution to the debt crises of the 1980s are reviewed in Adelman (1997). As Adelman notes, the economic history of the first half of the twentieth century was one of setback and growing difficulties. The momentum of world economic growth slowed significantly after 1913 and exhibited great disparities in rates of growth among world regions when compared to the decades immediately before and after this turbulent period. The first half of the twentieth century also witnessed a marked slowdown in international trade and an unprecedented depression from which the world struggled to recover. Two great wars, global in their scope, also describe this period.

It must also be understood that world economic possibilities at the close of hostilities in 1945 had been reduced by wartime devastation, and the fabric of international trade had been demolished by many years of intense fighting on land and sea and in the air. In many European and Asian countries, productive capacity in agriculture and industry had suffered severely, and in some of them, their prewar wealth, physical infrastructure, and channels of commerce had been completely destroyed. Even in those countries that had escaped the immediate effects of hostilities, economic activity and employment patterns had been seriously distorted by the demands of military expenditures and of a total war effort. Large imbalances described world patterns of production, prices, and international payments, and the economic imbalances were matched by political disorder and the emergence of a world divided by ideological conflicts. In this situation, it was by no means clear that the world economy could recover or, if it did, that a geographically balanced pattern of world development would emerge.

6. The long-term rise in world economic activity is usually measured by the growth in gross world product, a measure of the increase in the volume of goods and services produced across the globe during the period under review. There are numerous difficulties involved in constructing such an index, both practical and conceptual. Reliable data are, of course, an essential input. But the data must also be appropriately valued and converted into a common currency, and they must conform to certain theoretical principles. As one example, for proper valuation, prices used as weights in aggregating component series at the national level must represent on the margin the rate at which products of one series (e.g., consumption) can be transformed into the products of another series (e.g., capital formation); simultaneously, it must also measure the rate at which the production of one country can be transformed into the production of another. Needless to say, these assumptions are not and cannot be met in practice. Furthermore, price scales may change markedly over time in terms of their structure, and the application of one set of relative prices will yield different results than the application of another set.

The efforts of the United Nations to develop international statistical standards and measure quantitatively the state of the world economy and the changes taking place in its demographic, economic, and social development are reviewed in Ward (2004). Estimates of trends and conditions in the world economy cited here are based on those reported by international agencies and further standardized for country comparability as described in Walker (2008). As above, because of the many methodological and statistical deficiencies in the compilation of data relating to gross world product and international trade, any numbers cited should be seen as providing a general idea of levels and trends rather than precise figures.

7. The importance of economic growth in the lives of the world's poor is stressed by Easterly (2001).

8. The effects of population change on the growth of output are complex and not well understood, so it is difficult to say whether the higher rate of population increase in the later period affected the underlying rate of world economic growth. In the first instance, a rising population widens the market and allows for a greater division of labor and economies of scale. It may also stimulate technological change and the introduction of a wider variety of products. This would tend to raise the potential for economic growth. On the other hand, faster population growth raises the dependency ratio, lowers saving at any given level of per capita income, and necessitates capital widening (as the available capital is spread over a greater number of workers) rather than capital deepening (accumulating more capital per worker). It also increases the demand for public services and has a deleterious impact on the environment. This would tend to lower the potential for growth. The point made here is simply that the overall pace of world economic growth after 1950 would be lifted, ceteris paribus, by a faster rate of increase in population. This difference is on the order of 1% a year.

9. The Bretton Woods international monetary system was originally designed to address four problems that had arisen during the interwar years and were seen as major constraints on the possibilities for a stable expansion of world trade once the war ended: a shortage of monetary gold, which was overcome in the International Monetary Fund (IMF) agreement by substituting holding of national currencies convertible into gold; the need to adjust fixed exchange rates in an orderly fashion, which was now to be made by international agreement at the IMF; the need to deal with extraordinary fluctuations in international capital flows, which was to be remedied in the IMF agreement by allowing countries to introduce capital controls in accordance with IMF rules; and the need for sharing the responsibility for balance-of-payments adjustment between surplus and deficit countries, which was addressed by allowing members of the IMF to discriminate against countries whose currencies are scarce on foreign exchange markets.

The very success of the Bretton Woods system in the immediate postwar years lessened the importance of many of these concerns, and by its very success, it brought about another set of problems beyond its ability to solve. The growth in the use of the U.S. dollar and other national currencies as international reserve currencies, for example, solved the problem of a shortage of monetary gold but led to chronic and substantial deficits on current accounts in the United States and elsewhere. This and other problems, in turn, engendered a loss of confidence in the key currencies of the system, a growing liquidity problem as holdings of reserve currencies grew faster than the reserves of gold backing these currencies, and fundamental and necessary adjustments to balance-of-payments deficits failed to be undertaken or were avoided for domestic policy reasons. See Johnson (1965) for a discussion of the problems of monetary reform in the 1960s.

The general weakening of the system over time and the failure of international monetary cooperation to reform the system so as to control the increase and distribution of international reserves and support the process of balance-of-payments adjustment are only

the proximate reasons for the collapse of the system. More fundamentally, the very purposes of the international monetary system had been increasingly called into question. At Bretton Woods, the purpose of the monetary system was seen as promoting greater international trade in goods and services and rapid economic growth by releasing competitive forces and by providing a monetary system that would allow prices and costs to guide the process of world development. Less than three decades later, the purpose of the monetary system was increasingly seen in the most economically advanced countries as a means to protect countries from the need to adjust as they pursued full employment and in the developing countries as a means to promote domestic development in accordance with domestic priorities.

With this change in purpose came a marked slowdown in the pace of world economic growth and growing disparities in the performance of its different regions. It also left the world at the end of the twentieth century with an international economic system inferior to the system that prevailed at the start of the twentieth century.

10. The concept of productivity, as well as its measurement and studies of trends in productivity, is discussed in Helpman (2004, chap. 3).

11. A main reason for the faster rise in productivity in the commodity-producing sectors has been the difficulty of applying new inventions and new knowledge in labor-intensive service activities. Investments and innovations directed at increasing productive capacity require many adjustments resulting from the introduction of new plant and equipment and new ways to produce and sell products, whether old or new. Increasing the level of output of standardized items such as those produced in the commodity-producing sectors requires less investment and results in greater economies from long production runs than appear possible in services, where substitutes for labor are more difficult to find and output must be tailored to the needs of particular customers. Consequently, productivity gains from technological innovations and capital investment may be lower and costs associated with introducing change may be higher in services than in agriculture and industry. The change in economic structure during this period may be one reason why the potential for world economic growth may have fallen over time as production in all countries has become increasingly service oriented.

## References and Further Readings


Adelman, I. (1997). *The genesis of the current global economic system* (Working paper). Berkeley: University of California. Retrieved February 28, 2008, from http://are.berkeley.edu/~adelman/KEYNOTE.DOC

Easterlin, R. A. (1998). *Growth triumphant: The twenty-first century in historical perspective.* Ann Arbor: University of Michigan Press.

Easterly, W. (2001). *The elusive quest for growth: Economists' adventures and misadventures in the tropics.* Cambridge: MIT Press.

Helpman, E. (2004). *The mystery of economic growth.* Cambridge, MA: Belknap Press of Harvard University Press.

Henderson, D. (2001). *Anti-liberalism 2000: The rise of new millennium collectivism.* London: Institute for Economic Affairs.

Johnson, H. G. (1965). *The world economy at the crossroads: A survey of current problems of money, trade and development.* Oxford, UK: Clarendon.

Kindleberger, C. P. (1973). *The world in depression, 1929–1939.* Berkeley: University of California Press.

Maddison, A. (2001). *The world economy: A millennium perspective.* Paris: Organisation for Economic Co-operation and Development (OECD).

Schumpeter, J. A. (1950). *Capitalism, socialism and democracy* (3rd ed.). New York: Harper & Row.

Walker, D. O. (2008). *Historical estimates of world economic activity, population and labor force, 1950–2007* (Data 2008/Version 1, Regent University Working Paper). Retrieved December 13, 2008, from http://thedougwalkertdj.blogspot.com/2008/12/documentation-for-historical-estimates.html

Ward, M. (2004). *Quantifying the world: UN ideas and statistics.* Bloomington: Indiana University Press.

Yeager, L. B. (1966). *International monetary relations.* New York: Harper & Row.

# 45

# International Finance

Steven L. Husted
*University of Pittsburgh*

International finance is the branch of international economics that deals with a variety of issues that are related to macroeconomic behavior, such as the determination of real income and the allocation of consumption over time, in a country that engages in international trade. As such, this field of study is also often referred to as international macroeconomics. Specific topics included in international finance include the balance of payments, exchange rate determination, and macroeconomic policy in an open economy. At the heart of this field is the fact that international economic activity between and among nations seldom, if ever, is perfectly balanced. Therefore, financial resources will tend to flow across borders. The core of international finance is the study of these flows, the markets where this activity takes place, the impacts these flows have on economic behavior, and the interaction between these flows and economic policy.

The purpose of this chapter is to provide an overview of the field of international finance. The focus of the discussion will be on the major topics studied in the field and on our current state of knowledge about these topics.

## Exchange Rates and the Foreign Exchange Market

The foreign exchange market is the market where domestic money can be exchanged for foreign, and hence it is where the prices of many currencies are set. The price of foreign money is known as the exchange rate. Unlike the New York Stock Exchange, the foreign exchange market is not located in any one location. Rather, it is best thought of as a worldwide network of commercial banks linked together by sophisticated communications technology. Without doubt, it is the world's largest market; recent Bank for International Settlements (BIS) estimates place the volume of *daily* worldwide trade at $3,200,000,000,000 (Bank for International Settlements, 2007). It is also one of the most efficient. It is characterized by low barriers to entry; a homogeneous commodity (money); many buyers and sellers, none of whom has significant market power; and almost perfect information. Thus, it possesses all of the characteristics of a perfectly competitive market. And because of its overall size and international dimensions, it is totally unregulated by any national or international government.

Most of the trading takes place in a small set of currencies, including the U.S. dollar, the euro, the Japanese yen, the British pound, the Swiss franc, the Australian dollar, the Canadian dollar, the Swedish krona, the Hong Kong dollar, and the Norwegian krone. Of these, the U.S. dollar is most often involved in trades. This is because it serves as a vehicle currency in the market. Exchange rates, the prices of currencies, are quoted in dollars around the world, and trades of two nondollar currencies are often handled via an intermediate purchase of dollars with one currency and then a sale of those dollars for the other.

The foreign exchange market is almost always open. Business hours overlap around the world. The major foreign exchange trading centers are in London, New York, and Tokyo, with London accounting for about a third of all activity. The market is busiest in the early morning London time when European and Asian banks are open and early morning New York time when New York and London banks are simultaneously open. Other significant centers of trading activity include Zurich, Singapore, and Hong Kong.

Major participants in the foreign exchange markets include commercial banks, investment banks, and other financial institutions. These institutions conduct transactions in a variety of financial instruments. Principal dealer banks have developed into an Interbank market, with more than 2,000 registered dealer institutions worldwide. Dealers trade with each other via the telephone or increasingly more often through electronic brokerage systems. The largest dealers trade in all developed economy currencies as well as a selection of developing country currencies. Some dealers act as market makers for certain currencies, quoting both bid and offer prices for these currencies and willing to use their own capital to ensure that transactions occur at the prices that they quote. Some banks, especially in several European countries, act as currency brokers. That is, they bring together buyers and sellers of currency, earning a commission in the process.

Nondealer customers in the Interbank market include corporations, fund managers, individuals, and central banks (i.e., national monetary authorities). Corporations buy and sell currencies to process international trade activity and to fund payrolls for foreign operations. Fund managers buy and sell currencies in the process of buying or selling financial assets. Many central banks intervene in the foreign exchange market by buying or selling their own currencies to influence currency values.

Table 45.1 contains a set of exchange rate quotes as of Monday, January 12, 2009. These rates represent quotes as

**Table 45.1** Selected Exchange Rates: Monday, January 12, 2009

| *Country/Currency* | *U.S. Dollar Equivalent Monday* | *U.S. Dollar Equivalent Friday* | *Currency per U.S. Dollar Monday* | *Currency per U.S. Dollar Friday* |
|---|---|---|---|---|
| Australian dollar | 0.6803 | 0.7024 | 1.4699 | 1.4237 |
| Brazil real | 0.4318 | 0.4441 | 2.3159 | 2.2517 |
| Canada dollar | 0.8228 | 0.8398 | 1.2154 | 1.1908 |
| 1 month forward | 0.8223 | 0.8392 | 1.2161 | 1.1916 |
| 3 months forward | 0.8223 | 0.8395 | 1.2161 | 1.1912 |
| 6 months forward | 0.8228 | 0.8402 | 1.2154 | 1.1902 |
| China yuan | 0.1463 | 0.1463 | 6.8366 | 6.8353 |
| Euro area euro | 1.3378 | 1.3431 | 0.7475 | 0.7445 |
| Hong Kong dollar | 0.129 | 0.1289 | 7.7548 | 7.7580 |
| India rupee | 0.02057 | 0.0208 | 48.6145 | 48.0769 |
| Japan yen | 0.011225 | 0.011082 | 89.09 | 90.24 |
| 1 month forward | 0.011228 | 0.01109 | 89.06 | 90.17 |
| 3 months forward | 0.011236 | 0.0111 | 89 | 90.09 |
| Mexico peso | 0.0725 | 0.0733 | 13.7874 | 13.6426 |
| Norway krone | 0.1428 | 0.1425 | 7.0028 | 7.0175 |
| Singapore dollar | 0.6714 | 0.6737 | 1.4894 | 1.4843 |
| South Africa rand | 0.0987 | 0.1022 | 10.1317 | 9.7847 |
| South Korea won | 0.000738 | 0.000742 | 1354.46 | 1348.44 |
| Sweden krona | 0.1245 | 0.1257 | 8.0321 | 7.9554 |
| U.K. pound | 1.4824 | 1.5168 | 0.6746 | 0.6593 |
| 1 month forward | 1.4808 | 1.5155 | 0.6753 | 0.6598 |
| 3 months forward | 1.4792 | 1.5138 | 0.676 | 0.6606 |
| 6 months forward | 1.4775 | 1.5124 | 0.6768 | 0.6612 |

SOURCE: *The Wall Street Journal* (www.wsj.com).

NOTE: Currency mid-rates based on trading among banks of $1 million and more, as quoted 4:00 p.m. Eastern time to Reuters and other sources.

of 4 p.m. on that day as well as 4 p.m. quotes from the previous Friday. The numbers opposite the country names represent *spot* exchange rates. These are prices for currencies to be delivered within two working days. The first two columns report these prices in U.S. dollars. The next two columns report these same prices denominated in local currencies. Column 3 (4) entries are reciprocals of column 1 (2). The rates that are quoted are the average between the *bid* price (the price dealers will pay for that currency at that particular moment in time) and the *offer* (or asked) price (the price that they are willing to sell that currency for, also at that moment). The difference between the two prices is known as the spread. For the most commonly traded currencies in Interbank trades, the spread is very small, typically less than .02% of the price of the currency. The spread represents a source of profit to currency traders. Extremely active competition between dealer banks for this business ensures that the widths of the spread on various currencies tend to be small. Widths increase the less often currencies are traded.

With more than 180 countries in the world, there are more than 16,000 possible exchange rates. Market practices allow traders to calculate all possible exchange rates based on prevailing dollar exchange rates. As already noted, the dollar typically serves as a vehicle currency for trades involving two other currencies; the exchange rate between any two nondollar currencies at any point in time is based on that pattern of trades. For instance, the exchange rate between the British pound and the euro is calculated using the following formula (Grabbe, 1996):

$$£/€ = £/\$ \times \$/€ .$$

The term on the left is the price of 1 euro in terms of pounds. It equals the price of 1 dollar in terms of pounds times the dollar price of 1 euro. The logic behind this formula is that it is based on the cost of using pounds to buy dollars and then using those dollars to buy euros.

For some countries, additional exchange rates are reported in the table below the spot rate. These rates are known as *forward* exchange rates. A forward rate is an exchange rate that is set today for a transaction involving a purchase or sale of foreign exchange at some future point in time.

Forward exchange rates are calculated using a formula known as *covered interest rate parity* (CIRP):

$$\frac{F}{E} \times (1 + i^*)$$

where $i$ equals the interest rate in the home country (in decimal units over the same time span as that for the forward rate), $i^*$ equals the foreign interest rate, $E$ equals the spot exchange rate (domestic currency per unit of foreign), and $F$ equals the forward rate (domestic currency per units of foreign).

The underlying theory behind this formula is that the forward market provides a way for investors who consider investing in foreign securities the opportunity to protect themselves against adverse exchange rate movements. If the securities in question have otherwise identical characteristics in terms of risk, liquidity, and tax treatment, then financial markets will move to equalize payoffs on these assets when measured in the same currency. That is,

$(1 + i)$ is the return on a home asset, and

$\frac{F}{E} \times (1 + i^*)$ is the covered return on the foreign asset.

Equating these two returns yields the CIRP pricing formula.

Exchange rates may be totally market determined, in which case that country is said to have a freely floating exchange rate. This is true of about 40 countries in the world, including almost all of the major developed economies. Market-determined exchange rates can be very volatile over the short run. The next section discusses the behavior of floating exchange rates in the short and long runs. In the rest of the world, countries adopt various policies aimed at limiting exchange rate movements. These range from fixed exchange rate policies, wherein a government announces a fixed price of one unit of a target currency such as the dollar or the euro in terms of its currency, to allowing exchange rates to move over time but in a limited fashion. Policies such as these require the governments to purchase or sell their currencies in the foreign exchange market on a regular basis to limit exchange rate movements. Factors that influence the choice of an exchange rate regime and the impact of the choice are discussed later.

## Exchange Rate Behavior in the Short and Long Runs

Market-determined exchange rates exhibit considerable volatility. A variety of studies shows that the volatility of short-run exchange rate returns is indistinguishable from stock or bond market return volatility. Nelson Mark (2001, Table 3.1) provides a recent example of such calculations. Because of this similarity, most economists rely on *asset market models* to explain short-run exchange rate behavior.

The chief characteristic of an asset market model is its emphasis on forward-looking behavior. Asset prices today are determined in large part on expectations of the future performance of an asset. If people think an asset will rise in value in the future, they will be willing to pay more for that asset today, and its price will tend to rise. The same logic holds for foreign currencies.

To introduce exchange rate expectations into a formal model of exchange rate pricing, begin with the pricing formula for the forward rate derived in the previous section and introduce speculative activity. Let $exp(E_{t+1})$ equal the expected value of the exchange rate at time $t + 1$ (e.g., one month from now), based on information known at time $t$ (e.g., today). That is, this is the value that speculators expect

will prevail in the future. The forward rate, $F$, is set *today* for transactions that will occur in the future. The speculator takes a position in forward exchange based on the value of $F$ (which she knows today) and her guess of the future spot rate, $exp(E_{t+1})$.

Under certain situations, including perfect capital mobility and risk neutrality, and if speculators share expectations and act as a group, there is a tendency for $F$ and $exp(E_{t+1})$ to equalize. In this situation, the foreign exchange market is said to exhibit *uncovered interest rate parity* (UIRP). UIRP is a common feature of asset market exchange rate models.

Formally, using CIRP and substituting $exp(E_{t+1})$ for $F$, UIRP can be written as

$$E = exp(E_{t+1}) \times \frac{(1+i^*)}{(1+i)}.$$

This equation provides the basic model for explaining short-run exchange rate behavior. It states that the value of today's exchange rate depends on several things: domestic and foreign financial conditions (signified by the interest rate terms) and expectations about future values of the exchange rate.

Note the role of expectations in this model. If $exp(E_{t+1})$ rises, then $E$ will also rise. What could cause expectations to change? Almost anything can affect expectations, but especially changes in actual or anticipated government policies. Market participants are constantly searching the news for information that may influence exchange rates. Thus, the upshot of UIRP is that exchange rates should be very volatile because new information is constantly reaching the market. As this "news" is processed, it translates into buy or sell orders, which in turn cause exchange rates to move up and down.

Tests of the theory of UIRP are difficult to implement because a necessary element of the test is data on market expectations. Such data are difficult if not impossible to obtain. A standard procedure is to assume that the market has *rational expectations* of future exchange rates. Under that assumption, realized future exchange rates can be shown to equal their expected value at time $t$, plus a forecast error that is independent of information known by the market at the time $t$. The test then often amounts to a linear regression between actual changes in the exchange rate between $t$ and $t + 1$ and the interest rate ratio. If the theory holds true, then the regression coefficient on the interest rate differential should equal 1. In practice, however, a common finding is that the coefficient is significantly less than 1 and often less than zero (MacDonald, 2007, chap. 15).

There is a variety of explanations for this finding. One possibility is that expectations may not be rational; some market participants, known as noise traders, may take positions based on extraneous information rather than market fundamentals (Mark, 2001, chap. 6). A second possibility is that it may take time for market participants to learn about the full structure of the economy, and while they are learning, they make systematic prediction errors. This phenomenon is known as the peso problem (Krasker, 1980).

A third explanation assumes that the forward exchange rate includes a risk premium. UIRP assumes that the two assets in the model have identical risk characteristics. Hence, because these two assets are identical except in terms of denomination of currency, the expected return on each should be identical in the minds of potential investors. Note, however, that it is seldom the case that assets issued in different countries will have identical risk characteristics. Rather, it is more likely that the market will perceive one or the other to be inherently more risky to hold. In that case, market participants will require an additional return, known as a risk premium, $rp$, to be willing to hold that asset. In this case, the exchange rate equation will become

$$E = (F + rp) \times \frac{(1+i^*)}{(1+i)}.$$

In this case, $exp(E_{t+1})$ equals $F + rp$. Unfortunately, ex ante risk premia are virtually impossible to measure in the real world. Moreover, they are likely to be highly volatile. Hence, when real-world concerns about risk are included in the model, the theory predicts that exchange rates will be even more volatile than simple UIRP suggests, and it also becomes much more difficult to test the theory.

Given the short-run volatility in exchange rates, is there anything we can say about their long-run movements, such as the average annual change over a decade? Perhaps surprisingly, the answer is yes. In particular, while economists expect exchange rates to fluctuate considerably, due to asset market conditions, in the long run, we expect that their movement will be anchored by goods market considerations. The long-run pattern we expect is known as purchasing power parity (PPP).

The notion of PPP is one of the oldest concepts in economics. It refers to the idea that the same basket of goods should cost the same when prices are measured in the same currency regardless of where it is located. So, for instance, suppose that $P$ is the price of a bundle of goods in the United States, and let $P^*$ equal the price of an identical bundle in Italy (measured, of course, in euros). If the two bundles are to have the same price, the following equation must hold:

$$P = E_{PPP} \times P^*$$

or, equivalently,

$$E_{PPP} = P/P^*.$$

The theory of PPP says that the long-run equilibrium value of the actual exchange rate will be $E_{PPP}$. According to the theory, at any point in time, $E$ will probably not equal $E_{PPP}$. This is because the foreign exchange market is very volatile, subject to sudden shifts in demand and supply in

response to changes in expectations and other financial market considerations. However, given enough time, the theory of PPP says that the actual exchange rate should converge toward $E_{PPP}$.

There have been a large number of empirical tests of PPP. Initial tests, using relatively short spans of data, were not supportive. This is unsurprising given the fact that the theory argues that PPP is a long-run and relatively weak equilibrium condition. Data using longer spans of data and/or panels of data with a number of countries over a common period of time have been more supportive (Diebold, Husted, & Rush, 1991; Kim, 1990).

The theory of PPP is a statement that exchange rates and domestic and foreign price levels should move together in the long run. It says nothing about what causes any of these three variables to move. To close the circle, we need to add elements to the model. This is done with a theory of exchange rate behavior known as the monetary approach to exchange rate determination (MAER). The MAER is the workhorse theory of long-run exchange rate behavior. It was developed in the 1970s by economists at the University of Chicago and has been widely studied over the past 40 years.

The MAER has two fundamental building blocks. The first is purchasing power parity. The second is that agents in the two countries in question have well-defined stable demands for real money balances as a function of national income and interest rates. Imposing money market equilibrium and PPP, it is straightforward to show that the theory predicts the following equation for the exchange rate:

$$E = M^S - M^{S*} + 1_1 (Y^* - Y) - 1_2 (i^* - i),$$

where the * denotes foreign variables, $E$ is the log of the exchange rate, $M^S$ is the log of the nominal money supply, $Y$ is the log of real national income, $i$ is the interest rate, $1_1$ is the sensitivity of money demand to changes in income, and $1_2$ is the sensitivity of money demand to changes in the interest rate.

The MAER equation spells out the basic predictions of the model. In particular, holding all other variables constant,

a. rise in the home (foreign) money supply will cause an increase (decrease) in the exchange rate (i.e., a depreciation [appreciation] of the local currency);
b. a rise in home (foreign) output will cause a decrease (increase) in the exchange rate (an appreciation [depreciation] of local currency); and
c. a rise in the home (foreign) interest rate will cause an increase (decrease) in the exchange rate (a depreciation [appreciation] of local currency).

Predictions in part (a) should seem straightforward. In essence, they say that if a country prints more of its own money (everything else held constant), it will decrease in value in foreign exchange markets. This is because a rise in home (foreign) money will introduce inflationary pressures in the home (foreign) economy.

Predictions (b) and (c) show how changes in variables that influence money demand (everything else held constant) also can influence the exchange rate. In particular, growth in home (foreign) income raises money demand and puts downward pressure on home (foreign) prices. Working through PPP, this lowers (raises) the exchange rate. Growth in the home (foreign) interest rate lowers money demand and raises home (foreign) prices. Working through PPP, this raises (lowers) the exchange rate.

Note prediction (c). The direction of change of the exchange rate to a change in interest rates is exactly the opposite of the change predicted by the short-run model of exchange rate behavior. Why do the two theories make opposite predictions? The answer has to do with the difference between the short run and the long run. In the earlier model, a rise in the home (foreign) interest rate is assumed to indicate a short-run tightening of monetary policy. As such, it reflects a reduction in the availability of home (foreign) money. And when home (foreign) money becomes scarcer, its value rises. In the MAER, a rise in the home (foreign) interest rate is assumed to reflect a rise in the expected rate of future inflation. In this scenario, money in the future is assumed to be worth less than it is today (because of a rise in anticipated inflation), and hence home (foreign) money loses value today.

As with tests of PPP, empirical tests of the MAER have produced mixed results. Early studies using only short periods of data were not supportive. More recent tests using longer spans of data and/or panels of data have been more supportive (Engel & West, 2005; Husted, 1999).

## Balance of Payments and Models of Current Account Behavior

The balance of payments (BOP) is a statement of all of the international transactions of a country over a certain period of time. Standard presentations of the BOP provide detailed information on four major types of international transactions: trade in goods, trade in services, trade in financial assets, and international exchanges of gifts.

The BOP accounts are maintained according to rules of double-entry bookkeeping. In principle, double-entry bookkeeping attempts to record each side of every transaction. It does so by identifying one side as a credit entry in the ledger and the other side as a debit entry. Moreover, because each side has the same value, total credits in the table must equal total debits. Unfortunately, and unlike what is commonly done in accounting class presentations, BOP accountants rarely, if ever, see both sides of international transactions. Instead, they take the data they have, such as the values of exports and imports, and enter each item in the appropriate line as a debit or a credit. And because some transactions are unrecorded and others are estimated by government officials, it is virtually

certain that total credits will not equal total debits, without a balancing entry.

What determines whether an item is a debit or a credit? In the BOP table for any given country, a *credit* is any transaction that leads to a payment of money by a foreigner to a resident of that country. A *debit* is any transaction that leads to a payment of money by a home country resident to a foreigner. Using these rules and the available data, government statisticians construct the BOP table. Once all of the items are entered, debit entries are totaled as are credit entries. If total credits (total debits) are larger, then a statistical discrepancy debit (credit) equal to the difference between the two totals is added to the table. Table 45.2 provides a simple BOP table for the mythical country of Guatland in 2010.

The pattern of entries in the table follows the convention that appears in most official publications. Debit entries in the table are denoted with minus signs; credit entries are unsigned. International trade in goods and services appears at the top of the table. Income inpayments (outpayments) represent earnings paid to domestic (foreign) factors of production (wages, profits, rents, interest) by foreign (domestic) firms. Net transfers represent net flows of gifts, charity, foreign aid, and other one-way payments. In this example, outflows exceeded inflows, and hence the entry carries a negative sign.

Capital flows denote net changes in financial holdings of that type of asset (or liability) over the year valued at market prices at the time the transaction occurred. Long-term flows refer to purchases or sales of assets whose remaining time to maturity exceeds one year. Examples of inflows include purchases of domestic firms by foreign firms, foreign purchases of domestically issued equity, and long-term loans from foreigners to domestic residents. Examples of outflows include domestic purchases of foreign firms or foreign equity as well as long-term loans made to foreigners. Other capital flows include short-term loans such as trade credit as well as international changes in holdings of bank deposits. If foreign money comes into the country in these forms, then those amounts would be credits in the BOP. Examples of debit items in this section of the table include domestic loans made to overseas companies to finance international trade or domestic funds deposited in foreign banks.

**Table 45.2** 2010 Guatland BOP ($ Billions)

| | |
|---|---|
| Merchandise exports | $1,200 |
| Merchandise imports | –$1,800 |
| *Trade balance* | *–$600* |
| Service exports | $700 |
| Service imports | –$400 |
| *Balance on goods and services* | *–$300* |
| Income inpayments | $250 |
| Income outpayments | –$225 |
| Net transfers | –$100 |
| *Current account balance* | *–$375* |
| Long-term capital outflows | –$300 |
| Long-term capital inflows | $150 |
| Other capital outflows | –$125 |
| Other capital inflows | $660 |
| Statistical discrepancy | –$50 |
| *Overall balance* | *–$40* |
| International reserves | $40 |

According to Table 45.2, during 2010, the value of imports exceeded exports of goods, but services exports exceeded services imports. Firms and individuals in Guatland purchased more long-term foreign assets than foreigners purchased of Guatland assets. But other financial capital inflows such as increases in foreign-owned deposits in Guatland banks exceeded Guatland activity in these types of transactions.

In addition to these activities, Table 45.2 also includes an entry for international reserves. Reserves are government-owned, short-term highly liquid assets than can be used by the government to settle international imbalances with other countries. Examples of reserves include gold as well as liquid government assets issued by major world economies, including short-term government bonds issued by the United States. If reserves are acquired (sold) by a government, then that is treated as an outpayment (inpayment) and is debited (credited) in the BOP. In the table, Guatland lost $40 billion in reserves during 2010. Note that the overall balance in this example equals –$40 billion. Thus, this balance provides a direct measure of the change in a country's international reserves.

Countries accumulate international reserves for many reasons. They can be used to help pay for imports in those periods when export earnings are low. Most commonly, they are used as a means to influence the value of their currencies in the foreign exchange market. If leaders of a country want to raise (lower) their currency's value, they buy (sell) it in the foreign exchange market with (for) international reserves. Thus, large values of the overall balance in a country's BOP table may be evidence of currency targeting behavior.

Much of international finance is concerned with the behavior detailed in the BOP. Primary focus is on the current account balance (CAB). The CAB equals the balance of goods, services, and income plus net transfers; in the example in the table, it equals –$375 billion. Because the number is negative, in this example, Guatland is said to have a CAB deficit of $375 billion. This number equals net difference between debits and credits of all items above that line in the table. Why is this number interesting? Consider the only things left out of the calculation—international financial

capital flows. Recall as well that in the table, total credits equal total debits. This means that everything not included in the current account balance must exactly cancel out with the current account balance total. Hence, in this example, Because Guatland has a deficit of $375 billion, the remaining entries must sum to +$375 billion. This means that there is a net surplus in assets trade of $375 billion or, equivalently, that Guatland made net sales of domestic assets to foreigners in the amount of $375 billion or, equivalently, experienced an increase in liabilities to the rest of the world of $375 billion. Conversely, if Guatland had had a current account surplus of, say, $40 billion in a given year, then it must have a net deficit in assets trade of $40 billion during that year. This would mean that during that year, the country had increased its net holdings of foreign assets by $40 billion, thereby seeing its international wealth position rise. The current account balance thus provides a measure of that year's change in a country's international wealth position.

There are two basic models of current account behavior: the elasticities model and the intertemporal model. The focus of the first is on the role that changes in the exchange rate might play in eliminating CAB imbalances. In particular, for most countries, imbalances in trade in goods and services form the vast bulk of CAB imbalances. In turn, trade in these goods depends in large part on their prices and on the income levels at home and in the rest of the world. The elasticities model provides a simple way of understanding whether exchange rate movements can affect the current account in the desired direction.

For instance, the elasticities model concludes that if the supply curves of imports and exports are infinitely elastic, a necessary condition for a devaluation to improve the current account balance is

$$\eta + \eta^* > 1,$$

where $\eta(\eta^*)$ is the home (foreign) price elasticity of demand for imports. This equation is known as the Marshall-Lerner condition. Extended versions of the model allow for upward-sloping export supply curves as well as changes in national income levels to influence trade flows. When these additional factors are added to the model, the Marshall-Lerner condition becomes sufficient to ensure improvement in the current account following a devaluation. Because of the importance of the size of trade elasticities in the model, much empirical work has been undertaken over the years to measure trade elasticities. In general, econometricians involved in these efforts have concluded that although price import elasticities tend to be small, the Marshall-Lerner condition appears to hold (Houthakker & Magee, 1969).

In contrast to the elasticities model, which focuses on a CAB imbalance as an imbalance between exports and imports, the intertemporal model concentrates on international flows of assets needed to finance CAB imbalances. These flows can be shown to be the difference between the desired levels of saving and investment of the country. Decisions involving both amounts saved and invested are fundamentally decisions involving the passage of time (hence the name *intertemporal*). A person who saves postpones current consumption to be able to consume (ideally more) at a later date. Decisions by firms to invest involve decisions to purchase plant and equipment to be able to produce more (or more efficiently) in the future. International trade in financial assets allows economies to borrow from the rest of the world to finance additional investment or to lend to the rest of the world by providing their excess savings to finance foreign investment projects.

The basic idea of the intertemporal model can be established using national income accounting identities combined and some simple assumptions about macroeconomic behavior. We assume that in each period, individuals make plans about how much they want to consume and invest. Planned consumption we denote as $C^d$, and planned investment spending we denote as $I^d$. For simplicity, we treat government spending as totally exogenous. In addition, we will treat the CAB as residual. Its value is determined as an outcome of the plans made for spending on goods, services, and capital goods.

Given this setup, goods market equilibrium is achieved when planned spending equals production:

$$Y = C^d + I^d + G + \text{CAB}$$

or, equivalently, when planned saving equals planned investment spending plus the CAB:

$$NS^d = I^d + \text{CAB}$$

(where $NS = Y - C - G$ stands for national savings). Rearranging this last equation yields the fundamental equation of the intertemporal model:

$$\text{CAB} = NS^d - I^d.$$

That is, the current account balance is simply the difference between the desired saving level of a country in a given period and the desired level of investment spending. If national saving exceeds investment, there will be a current account surplus and vice versa.

The model makes a number of important predictions. Suppose a small open economy with access to the world credit market begins at a zero current account balance, then

a. if desired investment rises (falls), the country will incur a current account deficit (surplus);
b. if desired savings rises (falls), the country will incur a current account surplus (deficit); and
c. if desired savings in the rest of the world rises (falls), then world interest rates will fall (rise) and the home country will experience a current account deficit (surplus).

The intertemporal model has become the preferred approach to studying CAB behavior. It allows for a rich set of factors, including national demographic characteristics, interest rates, productivity trends, and so forth, that may influence a country to save or invest more and thereby affect its current account balance.

## Open-Economy Macroeconomics

One of the major lines of study in international finance is on the roles that trade and international capital flows play on overall macroeconomic activity. The workhorse model to tackle these issues is known as the Mundell-Fleming (MF) model. The MF model extends the standard IS-LM model by adding, wherever appropriate, international components that affect equilibrium conditions in the goods and assets markets. It also includes an additional curve that, depending on the exchange rate regime in place for the country, identifies the equilibrium condition for a stationary value for the exchange rate or for a zero balance in the overall balance of payments. This curve is called the BB curve. Solving the overall model will involve deriving the behavior of each of the three curves and finding a point of common intersection that will determine the equilibrium values for each of the endogenous variables in the model.

The basic MF model has two versions. The first version considers a small, open economy with perfectly flexible exchange rates. In this version, the endogenous variables are national output, $Y$; the interest rate, $i$; and the exchange rate, $E$. These variables will be determined by market forces to achieve equilibrium in the goods and assets markets as well as a zero overall balance of payments.

A second version of the model considers a small, open economy with fixed exchange rates. In this version of the model, the exchange rate is set by the government and ceases to be an endogenous variable. Instead, the overall balance of payments is determined by macroeconomic activity, although in the long run, it must be the case that the overall balance of payments converges to zero.

As has become the case with the IS-LM model, much of the analysis using the MF model focuses on the role of economic policy in stabilizing a country's economy. As it turns out, even in its simplest form, the model makes sharp predictions about the effectiveness of policy. In turn, these predictions crucially depend on the degree of international capital mobility and on the exchange rate policy followed by the country in question.

Among the many predictions of the MF model, the small, open economy version suggests that

a. under fixed exchange rates with perfect capital mobility, monetary policy has no effect on the level of output or on interest rates, but fiscal policy is very powerful in changing the level of output. Similarly, other shocks to aggregate demand, such as autonomous changes in investment or consumption, also have large output effects.
b. under flexible exchange rates with perfect capital mobility, monetary policy becomes very powerful in affecting the level of economic activity, while fiscal policy loses all effectiveness.
c. under fixed exchange rates, a devaluation (revaluation) of home currency raises (lowers) domestic output.

The intuition behind many of these results is straightforward. Monetary policy loses power to affect output under fixed rates because it must be directed at keeping the exchange rate fixed. It gains power under flexible rates because changes in the money supply affect interest rates and exchange rates in reinforcing directions. Expansionary (contractionary) monetary policy lowers (raises) interest rates and depreciates (appreciates) home currency, leading to increased (decreased) spending on several components of output.

When the MF model is extended to examine large economies whose actions may affect economic activity in the rest of the world, additional predictions emerge:

- Expansionary monetary policy by a large country under flexible exchange rates has "beggar thy neighbor" characteristics. That is, output expansion in the home market coincides with output contraction overseas.
- Expansionary fiscal policy by a large country under flexible exchange rates has a locomotive effect; output expands at home and abroad.

The broad predictions of the MF model hold up well in empirical tests of the model and match events in recent economic history. Because of this and its relative simplicity, the model is used widely in policy settings. Nonetheless, criticism of the model has grown, and in the professional economics literature, it has largely been replaced by dynamic forward-looking models. These newer models are fully general equilibrium, have solid microeconomic foundations, and allow the researcher to undertake welfare evaluations of various policy initiatives. This line of research has become known as the "new open-economy macroeconomics." The leading paper in this field is by Obstfeld and Rogoff (1995). However, in most cases, these models are technically very complex, and they have yet to be widely adopted for policy use or incorporated into undergraduate-level textbooks.

## Exchange Rate Regimes and Regime Choice

Over the past 150 years, the world has seen two extended intervals when countries followed fixed exchange rate policies. The first of these was known as the gold standard; it prevailed from about 1870 to the start of World War I in 1914 and then was revived for a few years in the late 1920s. Individual countries set fixed prices of gold in terms of their currencies and then took actions to maintain

those prices. Because all of the currencies in the system were fixed to gold, they were also fixed to each other. Virtually every major country and many of the then developing countries in the world participated in this system, so that fixed exchange rates prevailed worldwide.

The second period of fixed exchange rates was known as the Bretton Woods system. The Bretton Woods era lasted from 1946 to early 1973. This system was administered in part by the International Monetary Fund (IMF). IMF member countries declared par values for the dollar in terms of their currencies and took steps to maintain those values. In the process, each country's currency was linked to all others in the system.

In both cases, the worldwide system of fixed rates collapsed as individual countries came under pressure to change their exchange rates. Depending on the country and the era, fixed exchange rate policies were replaced by a variety of alternative policies, including floating rates or managed floats. Some countries have continued to fix the value of their currency in terms of another. The international monetary system that has been in place since the collapse of Bretton Woods features this mix of policy choices.

Slightly more than half of the IMF members currently pursue some sort of fixed exchange rate policies; the central banks of these countries must actively intervene in the foreign exchange market to maintain the stability of their currency values. Moreover, most of the countries that have floating exchange rates manage this behavior. That is, these countries also intervene in the foreign exchange market to manage the degree of changes in rates. Only about 40 of the members allow their exchange rates to be market determined at virtually all times. Most of these countries have large developed economies (International Monetary Fund, 2007). The proportion of countries that fix at least to some degree has been virtually constant since 2001. Prior to 2001, there had been a trend away from soft pegs to hard pegs or floats. That trend has now disappeared, although a handful of countries move from one category to another each year.

The choice between fixed and floating rates is of considerable interest to international finance economists. A variety of studies have shown that overall macroeconomic performance is not significantly different under the two regimes. Inflation tends to be slightly lower in countries that fix. Growth in per capita gross domestic product is about the same.

A clear difference between the two is that countries with fixed rates have abandoned the use of monetary policy for stabilization purposes. Choosing to fix, especially in developing countries, often follows periods of persistent inflation, brought on by excessive monetary growth. Announcing a fixed-rate policy provides the central bank with credibility that may enable it to risk expectations of future inflation. This is clearly an important reason for some countries to choose to fix rates. Another important reason is related to international trade. A number of countries choose to fix to their primary trading partner to prevent exchange rate changes from affecting trade flows.

In contrast, flexible rates are chosen most often by developed countries with powerful, independent central banks. Floating allows these banks to conduct activist monetary policy, and as shown by the Mundell-Fleming model, floating rates enhance the effectiveness of monetary policy. In addition, some economists view floating rates as providing an economic shock absorber and an automatic mechanism for helping countries deal with economic imbalances that might otherwise arise. There is some, but limited, evidence to support this hypothesis.

## Conclusion

As this chapter has sought to demonstrate, international finance is a broad field of study that deals with a variety of topics. Most important of these include exchange rate determination and behavior, balance-of-payments behavior, and macroeconomic policy and performance in an open economy.

## References and Further Readings


Bank for International Settlements. (2007). *Triennial Central Bank Survey December 2007: Foreign exchange and derivatives market activity 2007.* Basel, Switzerland: Author.

Diebold, F. X., Husted, S., & Rush, M. (1991). Real exchange rates in the nineteenth century. *Journal of Political Economy, 99,* 1252–1271.

Engel, C., & West, K. D. (2005). Exchange rates and fundamentals. *Journal of Political Economy, 113,* 485–517.

Grabbe, J. O. (1996). *International financial markets* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Houthakker, H., & Magee, S. (1969). Income and price elasticities in world trade. *Review of Economics and Statistics, 51,* 111–125.

Husted, S. (1999). Exchange rates and fundamentals in the short and long runs. *Australian Economic Review, 32,* 180–184.

International Monetary Fund. (2007). *Review of exchange arrangements, restrictions, and controls.* Retrieved from http://www.imf.org/external/np/pp/2007/eng/112707.pdf

Kim, Y. (1990). Purchasing power parity in the long run: A cointegration approach. *Journal of Money, Credit, and Banking, 22,* 491–503.

Krasker, W. (1980). The "peso problem" in testing the efficiency of the forward exchange markets. *Journal of Monetary Economics, 6,* 269–276.

MacDonald, R. (2007). *Exchange rate economics: Theories and evidence.* London: Routledge.

Mark, N. C. (2001). *International macroeconomics and finance.* Malden, MA: Blackwell.

Obstfeld, M., & Rogoff, K. (1995). Exchange rate dynamics redux. *Journal of Political Economy, 103,* 624–660.

# 46

# The European Economic and Monetary Union

Leila Simona Talani
*King's College London*

This chapter addresses the issues relating to the past, present, and future of the European Monetary Union (EMU), focusing on the way in which the main socioeconomic sectors within the most important European Union (EU) member states have used the process of European monetary integration to enhance their competitive position not only in the European arena but also in the global context.

This chapter will study in a historical perspective the phases leading to the establishment of the currency union, the present working of the European Central Bank (ECB) and the Stability and Growth Pact (SGP), and the future of the EMU.

## What Is the European Economic and Monetary Union?

EMU shall be read as a major institutional undertaking whose economic outcomes are not only constrained but also defined by its institutional framework. This is to say that the economic consequences of EMU are specific to the way in which this particular currency union has been devised, and this, in turn, is the product of a unique historical process.

Although the member states of the EU had been debating about creating an area of exchange rate stability in Europe since the end of the 1960s, its effective implementation took place only 30 years later.

Scholars identify four stages in the development of the process of European monetary integration.

The first stage goes from the Treaty of Rome (1957) to the Werner Report (1970). During this phase, the debate over the establishment of a single currency in Europe was almost nonexistent thanks to the international currency stability guaranteed by the Bretton Woods agreement from the immediate postwar period. However, the difficulty for the U.S. dollar to keep its value vis-à-vis gold as required by the Bretton Wood system, especially between 1968 and 1969, threatened the common price system of the common agricultural policy, a main pillar of what was then the European Economic Community (EEC). In response to this enhanced international monetary instability, the EEC's heads of state and government set up a high-level group led by Pierre Werner, the Luxembourg prime minister at the time, to report on how the EMU could be achieved by 1980.

The second phase spans from the Werner Report to the establishment of the European Monetary System (EMS) in 1979. This phase was characterized by the attempt to implement the Werner plan. This provided for a three-stage process to achieve the EMU within 10 years and included the possibility of introducing a single currency. The first stage in the implementation of the Werner plan was represented by reducing the level of instability of the European exchange rates. An attempt was made to achieve this aim through the establishment of the so-called Snake, an exchange rate arrangement based on fixing bilaterally the exchange rates among the European currencies. However, the international economic instability, especially the oil crises of the 1970s, made similar futile attempts, and the Werner plan was abandoned.

The third stage in the development of the process of European monetary integration started with the launch of the exchange rate mechanism (ERM) of the EMS in 1979 and ended with the issue of the Delors plan in 1989.

The EMS was based on three elements. The first was the ERM, which provided for the pegging of the exchange rates of the currencies participating in it within predefined fluctuation bands (ranging from ±2.25% to ±6%). The second element was the introduction of the European Currency Unit (ECU), a virtual, basket currency made up of a certain amount of each currency participating in the EMS. The last element was the very short-term lending facilities (VSTLFs) guaranteeing the possibility for all the central banks participating in ERM to intervene in the currency markets when necessary. The EMS succeeded in keeping the level of the exchange rate stable and in allowing for the introduction of anti-inflationary policies in most of the EU member states. It therefore represented the starting point for the discussion leading to the implementation of a full monetary union in Europe (EMU).

At the request of the European leaders, in 1989, European Commission President Jacques Delors and the central bank governors of the EU member states produced the Delors Report on how to achieve a currency union by the end of the 1990s. This represented the beginning of the fourth phase in the process of European monetary integration, which ended with the introduction of the euro in 1999.

The Delors Report reproposed the approach of the Werner plan based on a three-stage preparatory period. The three stages were as follows:

1. 1990–1994: completion of the internal market, especially full liberalization of capital movements
2. 1994–1999: establishment of the European Monetary Institute, preparation for the ECB and the European System of Central Banks (ESCB), and achievement of economic convergence through the convergence criteria (see below)
3. 1999 onward: the beginning of the ECB, the fixing of exchange rates, and the introduction of the euro

The European leaders accepted the recommendations of the Delors Report. The new Treaty on European Union, which contained the provisions needed to implement the EMU, was agreed at the European Council held at Maastricht, the Netherlands, in December 1991.

On January 1, 1999, the euro was introduced alongside national currencies. Euro coins and banknotes were launched on January 1, 2002.

From an institutional point of view, the EMU was defined by the 1992 Maastricht Treaty (Treaty on European Union) and its protocols. The treaty specifies the objectives of the EMU, who is responsible for what, and what conditions member states were required to satisfy to be able to enter the euro area. These conditions are known as the "convergence criteria" (or "Maastricht criteria"). They include permanence in the "new" ERM (±15% bands) for at least 2 years; inflation rates no more than 1.5% higher than the average of the three member states with the lowest inflation rate; interest rates no more than 2% higher than the average of the three member states with the lowest interest rates; a debt to gross domestic product (GDP) ratio not exceeding 60%, subject to conditions; and, most important, a deficit to GDP ratio not exceeding 3% (TEU art. 104(c) and art. 109 (j)).

With the introduction of the euro, the independent ECB, created only for this purpose, became responsible for the monetary policy of the whole euro area. The ECB and the national central banks of the member states compose the Eurosystem.

The Governing Council of the ECB, which comprises the governors of the national central banks of the euro area member states and the members of the ECB's Executive Board, is the only body that has the capacity to make monetary policy decisions in the euro area. These decisions theoretically are taken totally independently from the influence of the national member states' governments, parliaments, or central banks and from any other EU institution.

The treaty defined the only goal of the ECB, which is to ensure price stability within the euro area. According to the statute of the ECB, price inflation in the euro area must be kept below but close to 2% over the medium term.

National governments retain full responsibility for their own structural policies (notably, labor, pension, and capital markets) but agree to their coordination with the aim of achieving the common goals of stability, growth, and employment.

Also, fiscal policy (tax and spending) remains in the hands of individual national governments. However, they have agreed on a common approach to public finances enshrined in the Stability and Growth Pact (SGP).

The economic viability and the political (or political economy) rationale of the Maastricht criteria have been the subject of a number of studies from the most various academic points of view, which is why the issue will not be tackled here. What is important to underline is, however, that their strict anti-inflationary aim was even strengthened by the adoption of the SGP (Talani & Casey, 2008). The first version of the pact confirmed the objective of a deficit to GDP ratio not exceeding 3% and commits EMU member states to a medium-term budgetary stance close to balance or in surplus (Talani & Casey, 2008). It also defined the terms and sanctions of the excessive deficit procedure (EDP; Gros & Thygesen, 1998, p. 341).

What are the macroeconomic consequences of this institutional setting? How did the EMU affect the real economy of the member states? Did the EMU lead to more or less unemployment? The next section will try to answer these questions.

## The EMU and Unemployment

There are two ways of verifying whether the EMU increased or decreased the level of unemployment in the euro area. The first way is to assess to what extent this

particular form of currency union contains a recessive bias, thus reducing the level of output, ceteris paribus. The other way is to see how the establishment of the EMU has been linked in theory and in practice to the flexibility of labor markets. These two streams of reasoning might have opposite outcomes. Indeed, while the former would point to an increase in the level of unemployment, the second could lead to its decrease. However, the tricky aspect lies in the fact that the first way of reasoning might be used to further the second, thus adding to its economic rationale and fostering its political feasibility.

The first stream of economic reasoning assesses the overall recessive effects of the implementation of the Maastricht fiscal criteria and/or of the SGP. Some authors have argued that the effort brought about by the implementation of the Maastricht criteria, particularly the fiscal ones, as well as the determination to stick to the ERM in a period of high interest rate policy can explain the upsurge of European unemployment in the 1990s (Artis, 1998). Of course, economic analyses are far from reaching an agreement on the issue. Indeed, the counterarguments tend to underline the necessity of fiscal consolidation and anti-inflationary policies. Others point out that the time period over which unemployment has been growing in Europe is too long to be easily explained in macroeconomic terms (Nickell, 1997). However, deflationary policies implicit in the implementation of the Maastricht criteria to the EMU seem to have eventually worsened the level of European unemployment, at least by increasing the equilibrium rate of unemployment (a phenomenon called hysteresis in the literature; Artis, 1998; Cameron, 1997, 1998, 1999).

Moreover, some econometric simulations show that the implementation of the stability pact from 1974 to 1995 in four European countries (notably, Italy, France, Germany, and the United Kingdom) would have limited economic growth by reducing the annual growth rate (Eichengreen & Wyplosz, 1998, p. 93). This would lead to cumulated output losses of around 5% in France and the United Kingdom and 9% in Italy. The economic theory rationale of these results is, of course, that the SGP constraints would limit (and will limit, in the future) the use of automatic stabilizers to counter recessive waves, thus increasing the severity of recessions. This, however, will happen only if member states are not able to achieve a balanced or surplus budgetary position allowing them to use automatic stabilizers during mild recessive periods in the appropriate way without breaching the Maastricht/SGP threshold.

There is the counterargument that the stability and growth pact gives credibility to the ECB anti-inflationary stances, thus reducing the level of interest rates required to maintain the inflation rate below 2% and boosting the economy. Finally, even if the recessive bias of the fiscal criteria and the SGP were proved, this would not necessarily lead to a higher unemployment level (Artis & Winkler, 1997; Buti, Franco, & Ongena, 1997).

However, there is another way in which the Maastricht criteria and the stability pact might affect unemployment, a more indirect way, which is the basis to justify a neofunctionalist automatic spillover leading from the EMU to labor market flexibility. This is related to how member states should react to possibly arising asymmetric shocks. By definition, autonomous monetary policy and exchange rate policies cannot be used to react to idiosyncratic shocks in a currency union. At the same time, common monetary and exchange rate policy should be used with caution because it can have mixed results in case the other members of the EU are experiencing an opposite business cycle situation. Thus, economic theory leaves few options: fiscal policy, labor mobility, and relative price flexibility.

Indeed, a country could react to an asymmetric shock by using national fiscal policy, both as a countercyclical tool, through the action of automatic stabilizers, and in the form of fiscal transfers to solve more long-term economic disparities (as in the case of Italian Mezzogiorno). However, in the special kind of monetary union analyzed in this chapter, the Maastricht criteria and, to an even bigger extent, the requirements of the SGP constrain substantially the ability of member states to resort to national fiscal policy to tackle asymmetric shocks.

Alternatively, some authors suggest that the redistributive and stabilizing functions of fiscal policy be performed at the European level. Proposals include increasing the size of the European budget, pooling national fiscal policies, and establishing a common fiscal body, which would act as a counterbalance to the ECB (Obstfeld & Peri, 1998). The feasibility of similar proposals looks at least dubious in light of the difficulties EU member states encounter in finding some agreement on the much less challenging task of tax harmonization (Overbeek, 2000). Moreover, discussion about fiscal policy inevitably triggers a discussion on the loss of national sovereignty and a related one on political unification, whose outcomes are still far from being unanimous. Overall, there does not seem to be a compelling will of EU member states to reach an agreement on the creation of a common fiscal policy or to find some way to increase the size of the EU budget so as to introduce a stabilization function.

Given the difficulties in using national fiscal policy to tackle asymmetric shocks and the lack of any substantial fiscal power at the European level, economists suggest the option of resorting to labor mobility. The EU does indeed provide an institutional framework in which labor mobility should be enhanced. The treaty's articles regarding the free movement of workers, the single-market program, and the provisions about migration of course represent this. However, economic analyses show little evidence of mass migration in response to asymmetric shocks in the EU (unlike in some respects the United States; Obstfeld & Peri, 1998). Indeed, few European policy makers, if any, would seriously endorse temporary mass migration as a

credible way to react to national economic strains, for obvious political as well as social considerations.

There thus remains only one policy option for national policy makers to tackle the problems arising from asymmetric shocks: increasing the flexibility of labor markets so that "regions or states affected by adverse shocks can recover by cutting wages, reducing relative prices and taking market shares from the others" (Blanchard, 1998, p. 249). In addition, because reform of the labor market is clearly a structural intervention, it also will help eliminate the structural component of unemployment, apart from the cyclical one, if it is still possible to distinguish between the two (Artis, 1998).

Indeed, the employment rhetoric and strategy officially adopted by EU institutions in the past few years show clearly that the EU has chosen to give priority to labor flexibility and structural reforms as the means to tackle the problem of unemployment in Europe (Talani, 2008, chap. 8). However, the implementation of structural reforms is a costly endeavor that produces social unrest and requires the cooperation of domestic constituencies and actors. Is this effort going to disrupt the EMU and put strains on the process of European integration as a whole? What is the future of the EMU? Before turning to these questions, it is worth analyzing the performance of the ECB in terms of its monetary policy making.

## The Monetary Policy of the ECB From Its Establishment

Many were the worries concerning the performance of the ECB at the eve of its establishment, given the unprecedented nature of its tasks for being responsible for the implementation of a European common monetary policy and the management of a European common currency during a lack of full political integration. Some of the issues concerned a lack of credibility of the ECB monetary stances, a lack of flexibility, and a need to increase its democratic accountability and ensure its independence from the governments of the member states. At the onset, it is worth noting that the ECB is the most independent of all central banks. Indeed, its independence is guaranteed by its very statute to ensure as its exclusive goal the achievement of monetary stability, defined as a level of inflation below 2%. Independence from political constraints, both national and supranational, created further preoccupation over the democratic deficit of the European institutional setting but allowed central bankers to concentrate theoretically only on monetary variables, leaving aside the performance of the real economic indicators—namely, growth and employment rates.

The reality of the first years of implementation of a single monetary policy in the euro area, however, demonstrates that monetary policy considerations have not been separated from the performance of the real economy.

According to the Centre for Economic Policy Research (CEPR, 2000), from its inception, the ECB displayed more flexibility than expected regarding asymmetries within the euro zone. This result was possible thanks to the adoption of a so-called two-pillar monetary strategy at the expense of transparency. Given the goal of price stability, defined as a harmonized index of consumer prices (HICP) between 0% and 2%, the two pillars of monetary policy are, on one hand, a money growth reference target and, on the other hand, a number of unspecified indicators including the exchange rates and asset prices.

The first issue to address is the importance attributed by the ECB to output growth in euro land (and in some member countries in particular) relative to inflation.

Since the establishment of EMU, in 1999, the economic outlook recorded a marked slowdown in all Organisation for Economic Co-operation and Development (OECD) countries for the first time since the 1970s. Whereas the Japanese economy had been in a recession for some time already, in 2001, the U.S. economy experienced the first substantial fall of the business cycle in a decade. Also, the euro zone, with a lag of some months with respect to the United States, slowed significantly in 2001 (see Figure 46.1).

It is important to realize that this had a major impact, especially on the most important European economies—namely, Italy, France, and Germany (see Figure 46.2).

Theoretically, as underlined in many speeches and documents (CEPR, 2002), the ECB would pay little attention to the short-run output developments to avoid the threat of losing credibility in its anti-inflationary stances in front of the financial markets.

Despite this, even with a superficial analysis, it is easy to notice that the 30-point interest rate cut to 3% on January 1, 1999, was associated with deflationary risks in the wake of the Asian crisis. Furthermore, the April 1999 cut to 2.50% coincided with declining output in important euro land members (notably Germany). Finally, the cut on September 17, 2001, in the minimum bid rate on the Eurosystem's main refinancing operation by 50 points to 3.75, clearly matches a similar decision taken by the U.S. Federal Reserve in the aftermath of the terrorist attacks on September 11, 2001, and their recessive consequences (ECB, *Monthly Bulletin,* various issues).

More sophisticated analyses clearly show that the ECB monetary policy, particularly the timing and frequency of interest rate changes, reflected the aim to engage in some output stabilization and not only to control prices, although leading central banks personalities constantly denied it (CEPR, 2002).

If the output level was never officially recognized as a point of reference in the monetary policy making of the ECB but was certainly taken into consideration, the opposite happened with monetary targets. Indeed, the "two-pillar" strategy theoretically rests on the prominence of the target

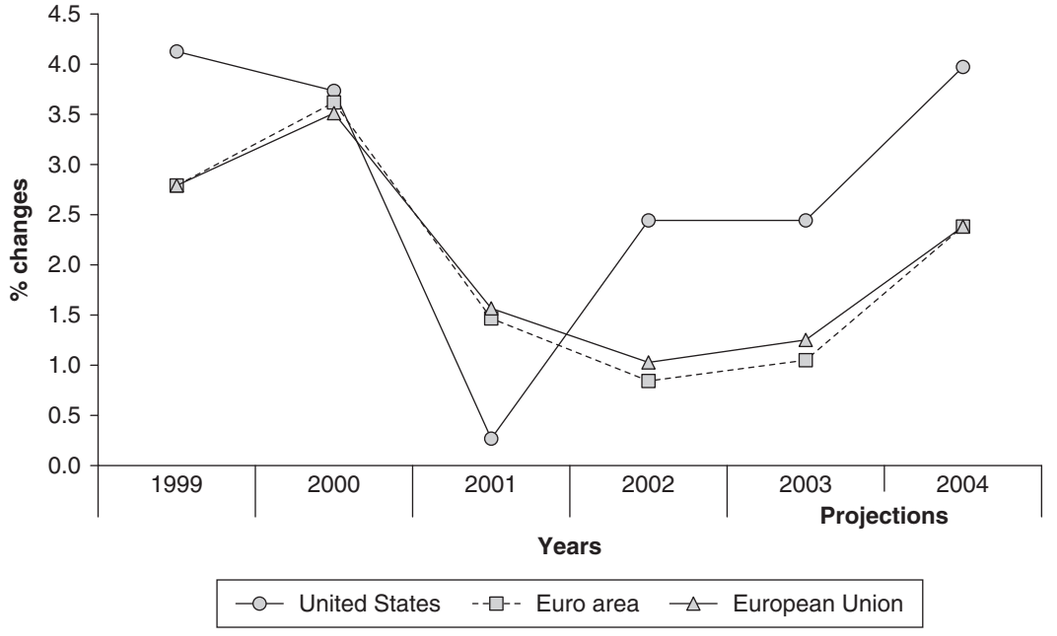


**Figure 46.1** Real Gross Domestic Product Percentage Changes, 1999–2004

SOURCE: Data from Organisation for Economic Co-operation and Development. Available from SourceOECD (http://www.sourceoecd.com).

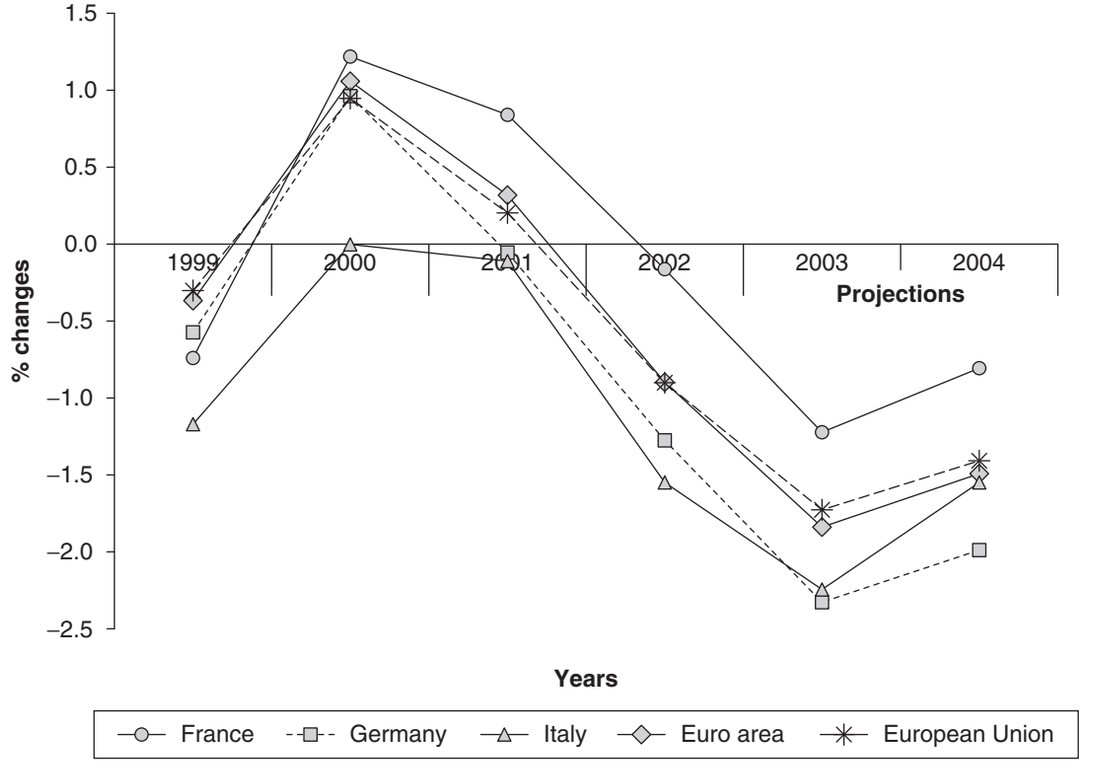


**Figure 46.2** Output Gaps

SOURCE: Data from Organisation for Economic Co-operation and Development. Available from SourceOECD (http://www.sourceoecd.com).

for M3 (monetary aggregate) growth as the main official indicator for ECB monetary policy decisions. However, on many occasions, when the target was overshot, the ECB did not react accordingly. For example, despite the fact that the target had been publicly set at 4.5% for 1999, no measures were taken by the ECB when it became clear that the target would not be achieved by the end of the year. On the contrary, the ECB cut the interest rates and engaged in sophisticated explanations about why the departure from the reference M3 growth rate did not represent any rupture with the two-pillar monetary strategy.

As the outlook for inflation turned upward by the end of 1999, the ECB did promptly intervene by increasing the interest rate by 50 basis points. Of course, given the parallel increase in the M3 growth, this seemed to be consistent with the monetary strategy declared by the ECB, while the final divorce between the ECB changes in the interest rates and the M3 growth rate appears justified by the necessity to keep the HICP within the 2% limit.

In any case, experts suspected that the M3 target was never really given the importance implicit in the adoption of the two-pillar strategy and was often subordinated to pragmatic considerations about the level of output. Indeed, reacting to the many criticisms toward the first pillar, the ECB effected some modifications of the M3 series by first removing nonresident holdings of money market funds from the definition of euro zone M3 and then purging nonresident holdings of liquid money, market paper, and securities.

However, this adjustment is no more than a cosmetic change and does not improve the reliability of the monetary pillar. If anything, Figure 46.3 shows that M3 percentage changes and interest rate decisions by the ECB, in its first years of activity, went in opposite directions.

Even more obscure is the role attributed by the ECB to the exchange rates within the two-pillar monetary strategy (CEPR, 2000). Indeed, the second pillar of the strategy makes explicit reference to a series of indicators influencing the ECB monetary decisions, among which are the exchange rates of the euro.

However, looking at the performance of the newly born currency in the first months of its existence raises spontaneously the suspicion that the bank had adopted an attitude of "benign neglect" vis-à-vis the exchange rate of the euro.

Indeed, the euro lost around 15% of its value vis-à-vis the dollar between August 1999 and August 2000, while the parity with the dollar was already lost in January 2000. Also, the effective nominal and real exchange rate of the euro experienced a marked decrease (–11.3 and –10.1, respectively, between August 1999 and August 2000; see Figure 46.4).

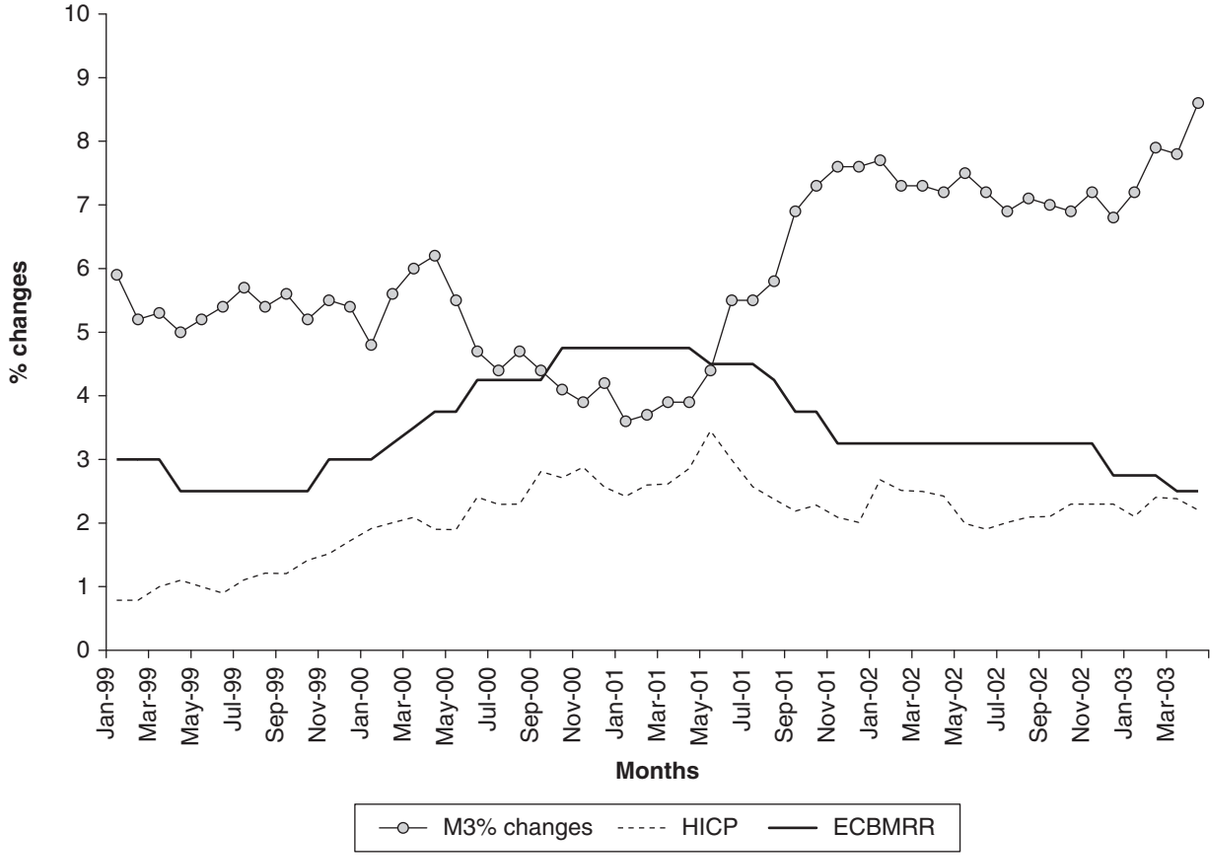


**Figure 46.3** M3, HICP, and ECB Main Refinancing Rate Percentage Changes, 1999–2003

SOURCE: Data from Organisation for Economic Co-operation and Development. Available from SourceOECD (http://www.sourceoecd.com).

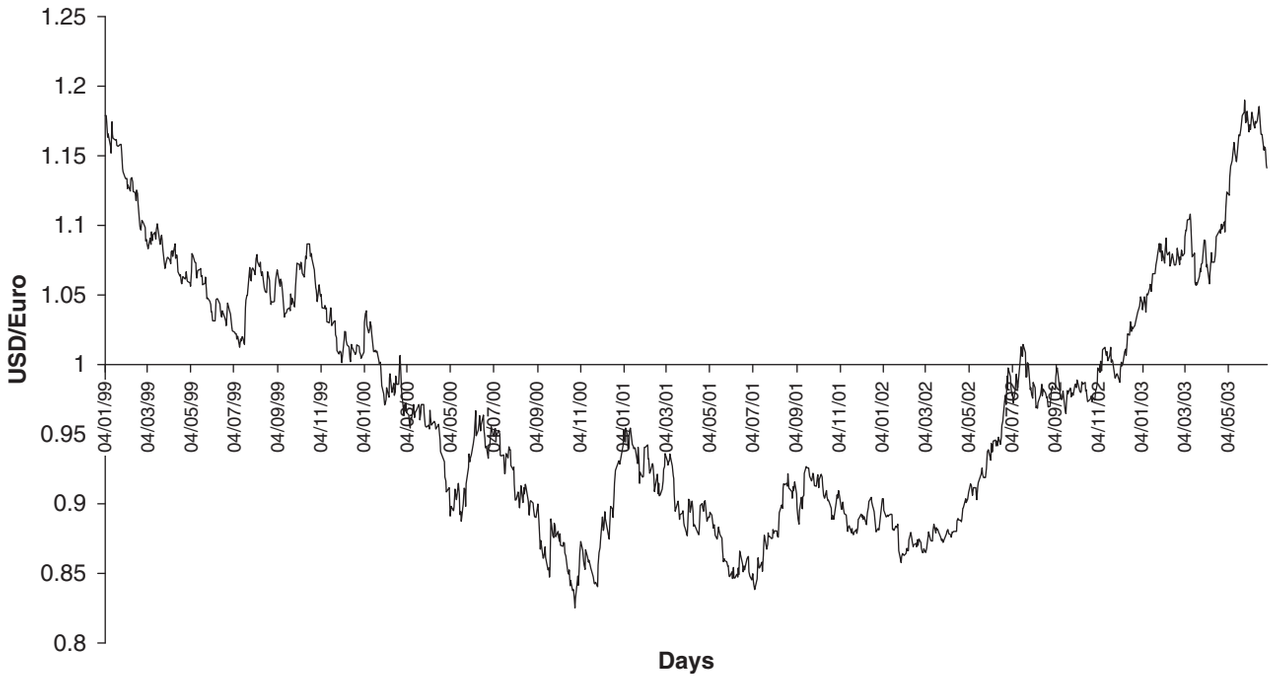


**Figure 46.4** U.S. Dollar/Euro Exchange Rates, 1999–2004

SOURCE: Data from European Central Bank (ECB). Available from http://www.ecb.int/stats/monetary/rates/html/index.en.html.

Of course, the ECB has always underlined that the performance of a currency must be assessed in the long run. And indeed, in the long run, the euro/dollar exchange rate has witnessed a reversal of its previous performance, with a marked appreciation of the euro, although it might be better to talk about a strong depreciation of the dollar.

However, the substantial lack of concern about the fall of the euro on the side of the European monetary authorities provoked further doubts about the real scope of the two-pillar strategy.

In a few words, the emphasis on the performance of the monetary aggregates (M3) as the first pillar of the ECB monetary strategy seems to conceal the desire by the central bank to trade off some of the transparency that the adoption of an alternative monetary strategy would imply (like targeting the inflation rate, for example), in exchange for more flexibility. In turn, this flexibility has been used to pursue output objectives that would not be acceptable otherwise within the strict anti-inflationary mandate of the ECB (CEPR, 2002).

Similarly, the attitude of the ECB toward the performance of the exchange rate—particularly in the first two years of its activity, an attitude that the economists fail to fully understand (Artis, 2003)—acquires a completely different meaning in light of analyzing the manufacturing export performance of the euro zone, particularly some of the euro zone countries (Talani, 2005). The export-oriented manufacturing sectors gained the most from a devalued currency.

Focusing on changes in the balance of trade in goods with the United States between 1995 and 2001, the data indicate that the countries recording the highest improvements of their trade balances with the United States were Italy (from 0.7% to 1.2%), France (from –0.3% to 0.55%) and Germany (from 0.6% to 1.5%) (Talani, 2005).

Therefore, the countries heavily relying on the performance of the export-oriented manufacturing sector, such as Italy, Germany, and France, had a vested interest in adopting a "laissez-faire" policy with respect to the depreciation of the euro.

Things, however, drastically changed when the dollar started depreciating, leaving members of the euro zone with the long-lasting problem of how to increase their economic competitiveness, especially after the globalization of the world economy started biting.

## The Future of EMU: Toward the Disruption of the European Monetary Integration Process?

What is the future of the EMU? Is the EMU unsustainable, especially in light of the global economic crisis? Will it lead to the disruption of the whole European integration process? Some answers about the disruptive potential of the EMU on European integration have already been given by the crisis and reform of the stability and growth pact.

*Refrigerated, hospitalized, dead:* These were the adjectives the press used to describe the SGP on the eve of the historic Council of Ministers for Economic and Financial Affairs (ECOFIN) decision not to impose sanctions on the delinquent French and German fiscal stances. The fate of the SGP was settled in the early hours on November 25, 2003, in what was a true institutional

crisis between the European Commission and the ECB on one side, demanding that the rules be applied, and the intergovernmentalist ensemble of the euro zone finance ministers rejecting the application of sanctions to the leading EU member states.

The most amazing thing was that, throughout the crisis, the euro remained stronger than ever and that the markets did not even think about speculating on the lack of credibility of a post-SGP EMU (Talani & Casey, 2008). Here a fundamental paradox arises: How can a currency remain as strong as ever in the midst of a serious crisis of the fiscal rule? The answer shall be sought in the interests of the socioeconomic sectors of the most important EU member states—namely, Germany and France. Here lies also the answer to the future of the EMU.

In other words, the economic interests of the French and German business sectors aimed at increasing their competitiveness, after relying briefly on the devaluation of the euro, with the reversal of this trend, focused on a reduction of taxes and on the implementation of the structural reforms (Crouch, 2002). In turn, the adoption of structural reform had still to rely on the consensus of the trade unions. This means that when the external, international conditions could not be modified by their previous institutional referents, such as the ECB, socioeconomic groups and the governments supporting their interests modified their policy preferences and decided to target other referents—in this case, ECOFIN. On the basis of similar considerations, it is possible to explain why, from 2002 onward, given the unlikelihood that the ECB could reverse or even slow down the depreciation of the dollar, the most powerful member states—namely, Germany and France—sought to obtain a relaxation of the macroeconomic policy framework, much needed by their economic domestic actors, by loosening the grip of the SGP. The exact timing of the crisis, in turn, was defined by the political needs of Germany involved precisely in November/December 2003 in the final stages of a tough negotiation with both the opposition and the trade unions for the approval of a package of structural reform denominated Agenda 2010 (Talani & Casey, 2008).

As underlined above, however, the demise of the SGP did not signify an abandonment of the EMU project but only a short-term contingent shift of the economic interests of the most powerful euro zone states. Therefore, the credibility of their commitment to the EMU remained intact, and the markets did not feel the need to attack the euro in the aftermath of abandoning the fiscal rule or to bet against the stability of the EMU by asking for higher yields. In brief, the future of the EMU was safe, the credibility of the EMU project was still rooted in structural considerations, and the decision to relax the fiscal rule was justified by a change of macroeconomic preferences by the leading socioeconomic groups in their quest for competitiveness. Concluding on a more general note, it is very likely that, despite the general economic crisis, the EMU will survive until it meets the economic interests of Germany and France.

## References and Further Readings


Artis, M. (1998). The unemployment problem. *Oxford Review of Economic Policy, 14*(3), 98–109.

Artis, M. (2003). *EMU: Four years on.* Florence, Italy: European University Institute.

Artis, M., & Winkler, B. (1997). *The stability pact: Safeguarding the credibility of the European Central Bank* (CEPR Discussion Paper No. 1688). London: Centre for Economic Policy Research.

Blanchard, O. J. (1998). Discussion to "Regional non-adjustment and fiscal policy." *Economic Policy, 26,* 249.

Buti, M., Franco, D., & Ongena, H. (1997). *Budgetary policies during recessions: Retrospective application of the stability and growth pact to the post-war period* (Economic Paper No. 121). Brussels: European Commission.

Cameron, D. (1997). Economic and monetary union: Underlying imperatives and third-stage dilemmas. *Journal of European Public Policy, 4,* 455–485.

Cameron, D. (1998). EMU after 1999: The implications and dilemmas of the third stage. *Columbia Journal of European Law, 4,* 425-446.

Cameron, D. (1999). *Unemployment in the new Europe: The contours of the problem* (RSC No. 99/35). Florence, Italy: European University Institute.

Centre for Economic Policy Research (CEPR). (2000). *One money, many countries: Monitoring the European Central Bank 2.* London: Author.

Centre for Economic Policy Research (CEPR). (2002). *Surviving the slow-down: Monitoring the EUROPEAN Central Bank 4.* London: Author.

Crouch, C. (2002). The euro, and labour markets and wage policies. In K. Dyson (Ed.), *European states and the euro.* Oxford, UK: Oxford University Press.

Eichengreen, B., & Wyplosz, C. (1998). The stability pact: More than a minor nuisance? *Economic Policy, 26,* 65–114.

Gros, D., & Thygesen, N. (1998). *European monetary integration.* London: Longman.

Nickell, S. (1997). Unemployment and labour market rigidities: Europe vs North America. *Journal of Economic Perspectives, 11,* 55–74.

Obstfeld, M., & Peri, G. (1998). Regional non-adjustment and fiscal policy. *Economic Policy, 26,* 205–247.

Overbeek, H. (Ed.). (2003). *The political economy of European unemployment.* London: Routledge.

Talani, L. S. (2005). The European Central Bank: Between growth and stability. *Comparative European Politics, 3,* 204–231.

Talani, L. S. (2008). The Maastricht way to the European employment strategy. In S. Baroncelli, C. Spagnolo, & L. S. Talani (Eds.), *Back to Maastricht: Obstacles to constitutional reform within the EU Treaty (1991–2007).* Cambridge, UK: Cambridge Scholars Publishing.

Talani, L. S., & Casey, B. (2008). *Between growth and stability: The demise and reform of the stability and growth pact.* London: Edward Elgar.

# 47

# East Asian Economies

Kiril Tochkov
*Texas Christian University*

Over the past 50 years, the economies of East Asia have attracted intense attention because of their rapid growth, which transformed them from relatively poor countries lacking modern technology to economic powerhouses with dynamic export-oriented industries and living standards similar to those in the richest countries of the Western world. Because of their rapid development, they have been referred to as "Asian miracles," which is justifiable especially if their transition is compared to the experiences of other countries. Initially, many East Asian economies were not very different from underdeveloped African countries in terms of gross domestic product (GDP) per capita, but their phenomenal growth enabled them to surpass the relatively wealthy South American economies and get close to the living standards in Western Europe and North America. Moreover, in only a few decades, East Asian economies experienced a development that took the United States and Western Europe more than 100 years.

Not surprisingly, economists have been analyzing the growth and development in East Asia and looking to identify the factors that might have contributed to this process. This is a very important and worthy exercise for two reasons. First, based on the analysis of the determinants of growth, it is possible to forecast whether current growth patterns of East Asian economies are sustainable in the long run. If the accumulation of physical capital through savings was responsible for their phenomenal growth, then their growth rates will probably slow down in the future. However, if their growth relied on technological innovation, it is likely that they will be able to sustain their rapid growth over the next few decades. Second, a detailed analysis of the East Asian growth experience can provide valuable lessons for other developing countries in terms of policy measures that stimulate growth.

The goal of this chapter is to provide a review of the various determinants of growth and development in East Asia in general and in China, Japan, South Korea, Taiwan, Hong Kong, and Singapore in particular. At first, the role of the two main factors of production, capital and labor, is examined. Next, the debate on the contributions of total factor productivity to growth is presented, followed by a discussion of additional growth components, including historical determinants, trade, and exchange rate policies. The last section deals with the origins and consequences of the East Asian financial crisis in 1997, which put a dent in the image of East Asian economies as "miracles."

## Demographics and Growth

To be able to produce goods and services, firms need to employ workers. Therefore, labor is one of the major determinants of aggregate output. Countries with large numbers of people who are able and willing to work (referred to as the labor force) usually also have a high GDP. China's population of more than 1.3 billion is the largest in the world, followed by India. Japan is also ranked among the top 10 most populous countries. Accordingly, China, with a GDP of more than $7 trillion in comparative terms, had the second largest economy in the world after the United States in 2007. India and Japan were ranked third and fourth, respectively. A high GDP provides countries with a larger pool of financial resources, which can be spent on national defense, allowing them to become regional or even world powers. However, GDP is not a good indicator for the standard of living in cross-country comparisons because it does not control for population size. For this reason, economists prefer to use GDP per capita, which reverses the rankings

based on the size of the economy as it rewards countries with a large GDP relative to their population. Accordingly, the tiny island of Singapore with a GDP per capita of around $50,000 in 2007 was ranked as one of the richest countries in the world, even richer than the United States. China, on the other hand, with a GDP per capita of around $5,000, was ranked at 100 given its large population.

Governments across East Asia recognized that to achieve prosperity, they would have to slow down the growth rate of the population. Family planning has been promoted as a national policy in several countries, including China, South Korea, and Taiwan. Since the late 1970s, the Chinese government has introduced policies that restrict the number of children to one per couple (hence the name "one-child policy") in urban areas and to two in rural areas, provide free contraception, and do not allow couples to marry before their early 20s. Those who violate these policies are subject to monetary fines, denied promotion at work, and harassed by government officials. In many cases, women who had given multiple births were forcibly sterilized or forced to undergo abortions. Although many of these measures are very intrusive and controversial, the policies have been largely successful in curtailing the population growth in China. In the medium run, this has proven very beneficial for the growth and development of China as it has eased the pressure to create jobs for the millions of people who join the labor force each year and has reduced the number of dependents that households have to provide for.

At the same time, population control measures have created imbalances that are likely to have an adverse effect on economic growth in the long run. One of the factors responsible for the emergence of dynamic export-oriented industries in the coastal areas of China is the existence of a large surplus of workers who have migrated from the rural areas and are willing to work longer hours for a relatively low wage. As the population growth slows down and the labor surplus is exhausted, the labor force will begin to shrink and the attractiveness of China as a low-cost destination for manufacturers is likely to diminish. Furthermore, the traditional preference for sons in China has created gender imbalances. As a consequence, millions of young men will not be able to find a wife in the future because of increased competition.

Japan, which is a much more mature market economy than China, is already experiencing some of these symptoms. It is one of the few countries in the world with a shrinking population, which is going to lead to shortages in the labor market in the future. Therefore, in contrast to the population control measures in China, the Japanese government has been trying to encourage families to have more children by providing financial incentives, building more day care centers, and offering women more generous maternal leave policies. Another possibility would be to allow more immigrants into the country, but such a policy has been highly unpopular in Japan. Interestingly, Japan has the highest robot density in the world and uses robots extensively in manufacturing, which might help reduce the negative effects of a shrinking labor force to a certain extent.

## Savings and Investment

Besides labor, firms need physical capital, including machines, tools, buildings, software and technology, natural resources, and land, to be able to produce goods and services. Physical capital is purchased by firms using financial capital, which in turn is borrowed in financial markets by selling bonds and shares and obtaining loans from banks and other financial institutions. The ultimate source of financial capital is the savings of the general population and of foreigners who are willing to purchase domestic financial assets. Consequently, savings and the resulting spending by firms on physical capital (called investment spending) are key determinants of economic growth.

Rich countries are able to accumulate large amounts of savings because of the higher level of aggregate income. In poor developing countries, where many people have to survive on less than a dollar a day, savings are insufficient, and firms have to rely on foreign capital, which either is borrowed if the country has good credit ratings or comes in the form of development aid. One of the distinctive features of East Asian economies is the high level of savings, which has helped them grow rapidly and achieve the status of "miracles." In the United States, consumption spending by individuals and households represents about 70% of aggregate spending in the economy, whereas the share of investment spending is only around 20%. In East Asia, consumption spending contributes less than 50%, while the share of investment spending is between 30% and 40% and can be as high as 50% in the case of Singapore. The main reason for these differences is the savings rate. Over the past 20 years, the personal savings rate in the United States has been very low and even negative in recent years. In East Asia, households save between 20% and 30% of their income. Given that the high savings rates are responsible for the rapid growth of East Asian economies, this phenomenon requires a more detailed analysis.

The research literature has provided several explanations (for China, see Modigliani & Cao, 2004; for Japan, see Horioka, 1990). People are able to save while they are young and able to work. As they grow older and their physical and mental abilities decline, they can retire and live off their savings. This means that the East Asian countries, with their relatively youthful demographic profile at the start of the growth period, were best positioned to achieve a higher savings rate. At the same time, this also means that the savings rate is likely to decrease in the future due to a fall in the birthrate (caused either by population control or by rising affluence) and aging, which change the demographic profile by increasing the share of retirees in

the population. A second important reason is the low level of social security in East Asian countries during the period of rapid growth. Although nowadays, Japan, South Korea, and Taiwan offer social security benefits similar to those in other rich nations around the world, these were introduced or amended only decades after their growth picked up. In China, the elderly have traditionally relied on their children and family for support. The socialist system in China has also ensured that those who worked in the cities and in state-owned enterprises enjoyed social security benefits. With the demise of the large family as a result of population control and the fundamental reform of the socialist economic system, people in China have been facing an uncertain future. A new system of social insurance was introduced only a few years ago, but it has a very limited coverage and is not functioning properly yet. The lack of insurance against unemployment or hardships after retirement has certainly encouraged people in East Asia to save more.

An additional but related issue was the lack of properly functioning financial markets. One of the reasons why Americans have a savings rate close to zero is because they have an easy access to credit due to the highly developed financial markets. In East Asia, this was not the case in the initial decades after growth accelerated. In China, mortgages and car loans were introduced only very recently, and financial markets remain largely underdeveloped. Under such circumstances where credit is unavailable, people are forced to save for years if they want to purchase durable goods or residential property.

Country-specific reasons also have contributed to higher savings and investment rates in East Asia. In Singapore, all employed individuals are required to contribute to the Central Providence Fund, a social security scheme run by the government (Peebles & Wilson, 2002). The contribution rates have varied between 40% and 50% of net wages in the past two decades and can be seen as a form of "forced savings." In China, the high investment rates are not only due to the abundant pool of domestic savings but have benefited from an enormous inflow of foreign direct investment. Since the 1980s, China has managed to attract Western firms with preferential conditions such as tax benefits, free land, and low wages. Since the 1990s, every major company in the world has set up shop in China, making it by far the largest recipient of foreign direct investment.

## The Role of Total Factor Productivity

Now that the two major factors responsible for the rapid growth in East Asia have been identified and discussed, it is important to examine their actual contributions to growth. As mentioned above, the pace of growth and development in East Asia has received considerable attention. From 1960 to 2000, the real GDP per capita in Japan, Taiwan, South Korea, Hong Kong, and Singapore grew at an annual rate of between 5% and 6% on average. China, which began a transition to a market economy in 1978, achieved an annual growth rate of almost 9% over the past three decades. In contrast, the corresponding growth rate in the United States and Western Europe from 1960 to 2000 was around 2%. This dramatic difference in growth rates suggests that East Asian countries will catch up with the Western countries and surpass them in the next few decades. However, for this to happen, East Asian economies would have to sustain the same high level of growth in the long run. Whether they are able to achieve this has sparked a major debate in the literature.

The particular issue at stake is the relative contribution of physical capital accumulation to growth in East Asia. According to the basic growth model developed by Robert Solow in 1956, economies grow as they accumulate physical capital through savings. For a given savings rate and population growth rate, the lower the amount of capital per worker, the higher the growth rate of output per worker. However, as the economy grows and continues to accumulate capital, the growth of output per worker slows down because of the diminishing marginal returns to capital. In other words, the model predicts that if East Asian growth was mainly fueled by investment in capital, then sooner or later the growth rates are going to decrease to the levels observed in mature economies of the United States and Western Europe. The Solow growth model offers two possible solutions to stem the slowdown. Countries could increase their savings rates, which would boost investment, or they could decrease population growth, which would contribute to higher levels of capital per worker. These measures are not suited for East Asians because of the extremely high savings rates in East Asia that are unlikely to rise further and because of the policies of population control that have already created problems such as an aging labor force. Furthermore, the two measures can boost the growth rate only temporarily. In the long run, the diminishing marginal returns to capital kick in.

When Solow tested his model using U.S. data, he found that besides capital and labor, a third factor also contributed to GDP growth. This third factor was not explained by the model and was termed *total factor productivity* (TFP). It is believed that TFP represents technological innovation and efficiency improvements. Given that theoretically there are no limits to technological and efficiency improvements, TFP becomes the only factor in the model that can result in sustainable growth in the long run. The debate on growth sustainability in East Asia thus focuses on how large the share of TFP is relative to that of physical capital accumulation.

The proponents of the idea that growth in East Asia is not sustainable at the high levels of previous decades have been dubbed the "accumulationists" because they believe that physical capital accumulation is by far the most important contributor to growth in these economies, whereas TFP plays only a minor role. The definitive work

from this viewpoint consists of a series of detailed empirical studies conducted by Alwyn Young and published in the mid-1990s (Young, 1994). The findings indicate that the growth of TFP in East Asia from 1966 to 1990 varied between 0.2% in Singapore and 2.3% in Hong Kong. In comparison, the growth rate of physical capital accumulation was shown to be between 8% and 14%. Although the TFP growth rates for East Asia are higher than for the United States or Western Europe, they are not large enough to explain the significant differences in growth rates of output per worker.

Inspired by Young's results, Paul Krugman, the recipient of the 2008 Nobel Prize in economics, published a famous article in 1994 that presented the accumulationist viewpoint in polemic terms. Krugman compared the development pattern in East Asia to that of the Soviet Union, which shocked the West with its claims of rapid industrialization and technological advances in the 1950s. At the time, many believed that the Soviet Union might be able to surpass the Western countries in terms of production and innovation. But in the following decades, it became obvious that the rapid growth of the Soviet economy was based entirely on increases in labor force participation, often using repressive methods, and on physical capital accumulation, achieved through forced savings. When the limits of manpower and the diminishing returns to capital took effect, the growth rate dropped, and in the absence of any technological innovation, the economy stagnated. The economic decline was ultimately one of the reasons for the breakdown of the Soviet system.

Krugman (1994) argued that the rise of the East Asian economies is very similar to the story of the Soviet Union in the 1950s, implying that because their growth was so dependent on physical capital accumulation and in the absence of efficiency improvements, it would eventually have to slow down. He focused on Singapore, which managed to double the share of the employed population, boosted its investment rate beyond 40% of GDP, and achieved significant improvements in the education level of the labor force. However, as Young (1994) had shown, TFP growth was zero, and because the limits to further increases in employment, investment, and education were reached, it was likely that economic growth would slow down significantly in the future. When the article was published in the early 1990s, East Asian economies were revered as economic miracles, and Krugman's attitude toward this prevalent view was expressed in his article's title, "The Myth of Asia's Miracle."

Krugman's (1994) article received wide attention in East Asia, particularly in Singapore, and provoked the publication of numerous articles that attempted to refute his arguments by adopting an "assimilationist" view of East Asian growth. Nelson and Pack (1999) agree that physical capital accumulation has been a major determinant of growth but argue that the technology that East Asian economies adopted from advanced countries was also assimilated in a manner that allowed genuine technological innovation to take place later. The process of assimilation encompassed entrepreneurship, learning, risk taking, creativity, and the adoption of better management techniques. Krugman did not see anything miraculous about the performance of East Asian economies because they mirrored similar attempts of other countries in the past that relied on investment to boost growth. In contrast, Nelson and Pack view the Asian economies as unique because of their efforts to learn and innovate from adopted foreign technology. To support their view, they used a sample of economies with very high investment rates and calculated the corresponding expected growth rate. Taiwan, Korea, Singapore, and Hong Kong were the only ones to significantly exceed the predicted growth rate, suggesting that additional factors must have played a role besides high rates of investment. If these factors reflected technological progress and innovation, then they would be able to sustain the extraordinary growth rates in East Asia in the long run.

## Additional Growth Determinants

So far, the discussion of factors that have contributed to the phenomenal growth of East Asian economies has focused mostly on the technical aspects of aggregate production. However, the growth performance of these countries was also affected by the environment in which growth occurred. Historical, political, geographical, cultural, and institutional factors have all played a role in shaping the economic behavior of individuals, firms, and the government in East Asia. Many of these variables are unique to the region and are thus crucial to understanding the rapid development that these countries have undergone.

### Historical Determinants

The modern history of East Asia begins with the arrival of European and American imperialism in the nineteenth century. At the time, China was considered to be the regional superpower. It was an empire as large as the whole of Europe with a history and culture going back thousands of years and surrounded by tributary states. The Western powers, including Great Britain, France, Germany, and the United States, were developing rapidly and were looking for trade opportunities around the world. East Asia was a lucrative market because it produced goods that were highly valued in the West, such as tea, silk, porcelain, and spices. The Chinese imperial government restricted trade with Europeans to the port of Canton. The Japanese imperial government, which had tried to isolate the country for centuries, allowed foreigners to trade only at the port of Nagasaki. To be able to buy Chinese goods, the British were exporting opium to China, which was grown in British India. On the other hand, the Chinese government was seeking to limit the import of opium. This conflict soon sparked a war that ended with the defeat of the Chinese and the signing of a treaty in 1842. China was required to pay large reparations, open several port cities to

trade, secede Hong Kong to Britain, and allow British citizens to reside and trade in China without being subject to its laws. In Japan, the United States also used "gunboat diplomacy" to sign similar treaties in the 1850s. Other Western powers soon followed the example of Britain and the United States.

The port cities that were opened to trade quickly became major commercial, financial, and industrial centers. This was facilitated by inflows of foreign capital and imports of advanced foreign technology. Shanghai, which was a small fishing village at the time, became a booming metropolis. Hong Kong, which was now a British colony, underwent a similar development. There is some debate about the impact of Western imports and production on the traditional industries in China. Some argue that cheap imports destroyed entire industries that relied on traditional technology and were not able to compete. Others suggest that there was a dual economy in which traditional and modern industries coexisted.

The humiliating defeat at the hand of the European powers weakened the imperial governments in China and Japan and caused numerous protests, riots, and rebellions. Government officials in both countries realized the need for fundamental reforms that would allow them to withstand the assault by Western powers. However, China and Japan went separate ways that were crucial for their future economic development.

Despite several reform proposals, the imperial government in China remained weak, indecisive, inefficient, and corrupt. It lacked a national policy of economic development and did not have the necessary capital to fund the industrialization of the country. In contrast, Japan strengthened the imperial powers and implemented a number of economic and political reforms that were vital for the development of the country in the twentieth century. Land and tax reform ensured a steady stream of revenue for the government, which it used to finance industrialization efforts. Modern factories were built, and foreign advisers were hired. Commercial banks were set up, and a central bank was created to oversee the monetary system. An efficient government administration was formed, and a constitution introduced elections and a parliament. Education became compulsory, and a modern army trained by foreigner military advisers was created. By the end of the nineteenth century, Japan had become a rapid-growing regional power that was starting to expand abroad. Soon it defeated China and Russia in regional wars and occupied Korea and Taiwan. Despite the destruction of its economy during World War II, Japan was able to recover quickly, drawing upon its experience of growth and development in the late nineteenth and early twentieth centuries.

## Political Determinants

The government played a crucial role during the period of rapid growth in East Asia, but very few countries in the region were democracies. Until the 1990s, South Korea and Taiwan were military dictatorships, and Hong Kong was a British colony. Singapore has been ruled by the same party since the 1950s. China has been a Communist country since the 1950s. This suggests that authoritarian political institutions are associated with economic growth in East Asia. This argument is supported by the findings of a seminal article by Barro (1996), who examined 100 countries over the period 1960–1990, which coincides exactly with the period of rapid growth in East Asia, and found that growth and democracy were negatively correlated.

While their economic policies have certainly contributed to rapid growth, authoritarian governments in East Asia have attempted to defend their repressive measures regarding human rights by creating a value system that claims to be better suited to the specific characteristics of the East Asian cultures than Western-style democracy. These "Asian values" have focused on the central role of an "enlightened" authoritarian government that knows what is best for the citizens of the country and achieves these goals using a mix of rewards and punishments. It promises political stability, social harmony, and economic growth, which are considered beneficial for the entire society. At the same time, citizens are expected to be loyal to and respectful of the government, avoiding demonstrations, criticism, and political debates. Those who dare to protest are severely punished.

Proponents of Asian values argue that democracy and human rights are Western concepts that are not suited for Asian cultures. Human rights are thus seen not as a universal concept that applies to all people across cultures. One of the most vocal supporters of Asian values is Lee Kuan Yew, a former prime minister of Singapore who ruled the island nation for decades and whose son is the current prime minister. However, not all East Asian countries are in favor of a system of Asian values that supports authoritarian governments. In the 1990s, South Korea and Taiwan made the transition from military dictatorships to dynamic democracies with free elections, multiparty systems, and respect for human rights and the rule of law. Former dissidents and democracy proponents who were jailed during authoritarian rule, such as Kim Dae-jung in South Korea and Chen Shui-bian in Taiwan, were elected presidents of their countries and strongly disagreed with the idea of Asian values providing cover for repressive authoritarian regimes.

## Trade and Exchange Rates

All East Asian economies have in common that their rapid growth was associated with an export-oriented development strategy. Government efforts focused on fostering industries that produced goods for the world market. In many cases, high-quality goods were produced exclusively for export, whereas more inferior goods were reserved for the domestic market. To generate profits from exports, companies had to learn about their competitors and their customers abroad and were forced to make improvements,

introduce new technologies, innovate, and increase efficiency to stay competitive. The government benefited as well by imposing export taxes and accumulating reserves of foreign currency.

Export-oriented industries received government support in several forms. They were provided with cheap credit, import tariffs were eliminated for inputs needed by these industries, labor unions with demands for higher wages and better working conditions were suppressed, and the exchange rate was devalued to make exports competitive. Initially, exports from East Asia consisted mostly of textiles, plastic, and steel, but over time these economies started producing more sophisticated products such as chips, computers, and cars. Lower production costs, competitive pricing, and a gradually improving quality and design enable East Asian exports to increase their market share in Europe and the United States. Over the 1970s and 1980s, trade frictions between Western and East Asian countries increased due to the growing trade surplus accumulated by East Asian economies, and calls for the protection of domestic industries in the United States became louder. To avoid trade sanctions, Japan, South Korea, and Hong Kong agreed to voluntarily restrain their exports to the United States and open up their markets for U.S. firms.

China was largely closed to the world until the late 1970s. Since then, export-oriented industries have sprung up along the coast, and China was transformed into one of the major producers of manufactured goods in the world. To attract foreign direct investment, the Chinese government created special economic zones that were located along the coast. Foreign companies were lured by beneficial conditions such as tax waivers. However, the main goal of the government was to enable domestic firms to learn from advanced technologies used in Western firms. For this reason, all foreign firms were required to team up with a Chinese partner. Over the 1990s, Chinese exports grew rapidly, and the large trade surpluses irritated the United States and the European Union. China joined the World Trade Organization in 2001 but has been accused by its trading partners of dumping products on foreign markets at below cost, restricting access to its domestic market for foreign firms, and not doing enough to protect intellectual property rights.

One of the major friction points concerns exchange rates. Most East Asian economies have had fixed exchange rates during the period of rapid economic growth. Fixed exchange rates provide stability but also give governments a tool with which they can manipulate trade flows. East Asian governments have tended to undervalue their currency. In other words, their currencies were exchanged for fewer U.S. dollars than would have been the case without a fixed rate. This is beneficial for the exporting industries because it makes Asian goods cheaper for the American consumer and boosts exports. Those who are hurt by this system are the U.S. exporters who have difficulties selling their goods to Asian consumers because of their relatively high price. The disadvantage of fixed exchange rates is that governments need to be ready to defend them against market fluctuations and speculative attacks. To maintain a fixed exchange rate, governments need foreign currency reserves that they either buy or sell on the foreign exchange markets.

China is one of the countries that have consistently kept an undervalued currency against the U.S. dollar. The resulting boom in Chinese exports to the United States has provided Chinese companies with U.S. dollars that they deposit in China. At the same time, the increased demand for the Chinese currency pushes up its value, making it necessary for the government to intervene and buy dollars. Over the years, China has accumulated more than $1 trillion in reserves. Given that not all the money is needed to defend the fixed exchange rate, the Chinese government set up an investment company that was given the task to buy shares in profitable projects around the world. In addition, Chinese companies, which are often partially owned by the government, have also been on the lookout to purchase foreign assets. Something similar happened in the 1980s when Japan was in the same position. At the time, Japanese companies bought famous U.S. companies and landmarks, such as Hollywood studios and the Rockefeller Center in Manhattan. The Chinese companies and the investment corporation have focused much of their attention on natural resources that China desperately needs to grow. Oil, natural gas, metal ores, and other minerals have attracted Chinese companies to invest in African countries, which are often considered too risky for Western companies. Other targets of Chinese investment abroad include companies with valuable brands or technology or an established market share. The most famous examples are the takeover of the IBM personal computer business by Lenovo, a Chinese computer firm, and the purchase of Rover, a British car producer, by a Chinese company. Last, Chinese investors have also poured money into U.S. government securities as well as in more risky shares in private equity firms.

As with the Japanese companies in the 1980s, the current inflows of Chinese investment have resulted in calls for protectionist measures against China, arguing that Chinese companies have an unfair advantage because of their undervalued currency and government subsidies, which they use to accumulate dollars that in turn are used to buy American assets. Furthermore, by purchasing U.S. assets, the Chinese pour money into the U.S. economy and thus contribute to cheap credit for U.S. firms and consumers. The resulting liquidity, however, has been blamed for the high debt levels among U.S. consumers, which ultimately caused the freezing of capital markets in 2008.

A possible solution to the problem of consistent trade surpluses and large accumulations of dollar reserves is for China to give up its fixed exchange rate. This would lead to the appreciation of the Chinese currency, making Chinese goods more expensive for American consumers and thus decreasing the trade surplus. At the same time, the stronger Chinese currency would increase the purchasing

power of the Chinese consumers who would now be able to afford more American goods. This would also reduce the U.S. trade deficit but would also rebalance the Chinese GDP, increasing the share of personal consumption at the expense of investment spending. The U.S. government has been trying for years to pressure China to allow its currency to appreciate. China has responded by fixing its currency against a basket of currencies rather than only against the U.S. dollar. In addition, the Chinese currency is allowed to fluctuate in a relatively narrow band. Despite these measures, the Chinese currency remains largely undervalued. The Chinese government is reluctant to let its currency float because export industries generate growth and employment. An appreciation of the Chinese currency could slow down growth and result in an increase in unemployment, which in turn could fuel social unrest.

## The Asian Financial Crisis

The Asian financial crisis is an important event in the development process of East Asian economies because it affected the entire region, plunged the economies into a severe recession, and shattered belief in the "Asian miracles." The crisis began in Thailand in 1997 and was associated with the large inflows of portfolio investment from abroad during the 1990s (Corsetti, Pesenti, & Roubini, 1999). Asian banks and financial institutions had borrowed heavily from abroad, which led to a boom in lending domestically. The main problem was that the short-term inflows from abroad were used for long-term lending, which made the financial system vulnerable. In addition, much of the domestic lending was used for risky projects and speculation on the property markets. When several financial institutions in Thailand missed their payments on foreign debt, international investors panicked and began pulling their money out of East Asia. This resulted in a wave of bankruptcies, closures, and layoffs. As in other Asian countries, the Thai currency was fixed to the U.S. dollar, which made it vulnerable to a speculative attack. The government attempted to defend the fixed exchange rate but, after losing billions of U.S. dollars in reserves in a matter of weeks, eventually gave up and allowed the value of the currency to be determined by the markets. The crisis soon spread across the region, and Singapore, Taiwan, and South Korea were forced to give up their currency pegs or devalue their currency to stay competitive.

China was one of the least affected economies because its financial markets were still underdeveloped, its currency was not convertible, and strict capital controls prevented inflows of portfolio investment from abroad. In contrast, South Korea was one of the most prominent victims of the crisis. Korean banks had borrowed abroad in foreign currency but loaned to the domestic conglomerates that dominate the Korean economy in domestic currency. When the Korean currency lost value, the foreign debt mounted, resulting in bank insolvencies. Several large conglomerates burdened with debt were also declared bankrupt. The Korean government was forced to seek financial help from the International Monetary Fund (IMF). The billion-dollar loan, however, was conditional on a deep restructuring of the financial sector, strict budget discipline, and further reforms of the industrial structure. Although Japan was also severely affected by the Asian financial crisis, it had experienced its worst recession in 1990–1991 when an asset bubble that had developed over the preceding years burst. What followed was a decade of extremely weak economic performance with growth rates of between 0% and 3%.

The slowing growth in Japan over the 1990s and the economic shocks of the Asian financial crisis tarnished the image of the "Asian miracle" and appeared to confirm the predictions by Krugman (1994). Although most East Asian economies managed to recover from the crisis a few years later, they were not the same dynamic economies of earlier decades. At the same time, China emerged as the new economic power in East Asia and continued to captivate the world with annual growth rates of more than 10%.

## Conclusion

From 1960 to 1990, several East Asian economies achieved phenomenal growth and became synonyms for successful development. This chapter provided an overview of the factors that have contributed to this extraordinary growth performance. Physical capital accumulation has been identified as the single most important determinant of growth and was largely the result of very high saving and investment rates, two common features across all East Asian economies. The danger associated with relying exclusively on investment is that sooner or later growth would slow down due to the diminishing marginal returns to capital. Measures aimed at population control have reduced the population growth rate and have boosted GDP per capita. However, they have also led to a more rapid aging of the population and thus risk having a negative effect on future growth. The share of TFP in growth has been a matter of debate, with some arguing that it was at par with other economies, while others insisted that it played a crucial role in the miraculous performance of East Asian economies.

Beyond the factors of production and technology, various other historical, political, cultural, and institutional factors have contributed to the development process. The first contacts with Western imperialism revealed the necessity of fundamental reforms to modernize the East Asian economies and societies. While Japan succeeded early on in learning from the West, China's inefficient government system, lack of reforms, and low revenue prevented it from taking an active role in the modernization of the country. Trade has also been a key factor in the success stories of East Asian countries, although it has increasingly led to frictions about trade surpluses, dumping, and subsidies with its trading partners. Fixed

exchange rates were also helpful to achieve growth, but in later decades, they became a liability and were eliminated. The Asian financial crisis was a severe blow for the economies involved and shattered the beliefs in the Asian miracle.

## References and Further Readings


Aghion, P., Alesina, A. F., & Trebbi, F. (2007). *Democracy, technology, and growth* (NBER Working Paper No. W13180). Cambridge, MA: National Bureau of Economic Research.

Barro, R. (1996). Democracy and growth. *Journal of Economic Growth, 1,* 1–27.

Bramall, C. (2009). *Chinese economic development.* New York: Routledge.

Corsetti, G., Pesenti, P., & Roubini, N. (1999). What caused the Asian currency and financial crisis? *Japan and the World Economy, 11,* 305–373.

Flath, D. (2005). *The Japanese economy.* New York: Oxford University Press.

Horioka, C. (1990). Why is Japan's household saving rate so high? A literature survey. *Journal of the Japanese and International Economies, 4,* 49–92.

Ito, T. (1991). *The Japanese economy.* Cambridge: MIT Press.

Krugman, P. (1994). The myth of Asia's miracle. *Foreign Affairs, 73,* 62–78.

Modigliani, F., & Cao, S. (2004). The Chinese saving puzzle and the life cycle hypothesis. *Journal of Economic Literature, 42,* 145–170.

Naughton, B. (2007). *The Chinese economy: Transitions and growth.* Cambridge: MIT Press.

Nelson, R., & Pack, H. (1999). The Asian miracle and modern growth theory. *The Economic Journal, 109,* 416–436.

Peebles, G., & Wilson, P. (2002). *Economic growth and development in Singapore: Past and future.* London: Edward Elgar.

Solow, R. (1956). A contribution to the theory of economic growth. *Quarterly Journal of Economics, 70,* 65–94.

Song, B. N. (2003). *The rise of the Korean economy.* New York: Cambridge University Press.

Young, A. (1994). *The tyranny of numbers: Confronting the statistical realities of the East Asian growth experience* (NBER Working Paper No. W4680). Cambridge, MA: National Bureau of Economic Research.

# 48

# Globalization and Inequality

Rachel McCulloch
*Brandeis University*

Goods and services now move more freely among countries than ever before. Ongoing declines in the cost of long-distance communication and transportation and in national restrictions on international trade and investment have allowed economies around the world to become increasingly integrated, thereby enhancing productivity growth and expanding consumer choices. In parts of the developing world and especially in East Asia, globalization has been accompanied by an increase in living standard hardly imagined just a generation ago. At the same time, globalization has also become the focus of widespread controversy. In particular, concerns about adverse consequences for income distribution have fueled policy initiatives that threaten to turn back the clock.

An especially troubling development was the emergence of a popular backlash to globalization in the United States, even when the country was enjoying record growth and the lowest unemployment rate in decades. An article in *The Economist* ("Globalization and the Rise of Inequality," January 18, 2007), written while the U.S. economy was expanding, highlights "a poisonous mix of inequality and sluggish wages" as the force underlying a globalization backlash in the United States as well as Japan and the European Union. Once America's long period of expansion reached an abrupt end, market-opening trade accords such as the North American Free Trade Agreement (NAFTA) became lightning rods for public concern about stagnating real incomes, job losses, and increased economic insecurity.

In the midst of the global recession that followed, most Americans and their counterparts abroad continued to acknowledge the benefits of globalization in terms of overall productive efficiency, lower prices, and increased consumer choices. The backlash, both in the United States and worldwide, has largely been a response to the perceived *redistributive* consequences of increased openness, especially openness to imports and immigration. Even if a country "as a whole" is made better off, and not everyone accepts this premise, there remain concerns about the well-being of particular groups within its borders. Indeed, increased globalization has been accompanied by increased inequality not only in the United States but in many other countries. But is globalization a major *cause* of increased inequality?

While there is little disagreement that globalization has been *accompanied by* increased inequality within many countries, both rich and poor, this is not the same as establishing a causal relationship. On the contrary, numerous systematic studies have concluded that the redistributive changes the public and policy makers often attribute to globalization are due mainly to other changes in the economy. Thus, while discussions of globalization frequently turn to its contribution to inequality, discussions of rising inequality often fail even to mention globalization or dismiss it as playing at most a minor role in explaining recent trends in income distribution. Many U.S. trade and labor economists conclude that the primary cause of increased wage inequality is that the rate of skill-biased technical change has exceeded the growth rate of skilled labor. However, most technical change is not exogenous but the result of profit-motivated investment. By altering the incentives firms face, openness to trade may promote development and adoption of new production methods. The rate of skill-biased technical change may also respond to increases in the availability of skilled labor, so that demand and supply interact over time.

Finally, even the fact of increased inequality is subject to challenge. While most studies confirm substantial inequality both within countries and between them, conclusions regarding recent trends are highly sensitive to specifics of

the methodology: definition of inequality, data sources, and estimation techniques. Do we mean greater wage inequality? For the United States, trends in household earnings tend to follow trends in wages because labor earnings are so important in total earnings. But in developing countries, a much larger share of income, especially at the low end of the distribution, comes from self-employment.

Although *globalization* typically refers to expanded trade in goods and services, much of the controversy actually arises from other aspects of global integration: immigration, foreign investment, technology transfer, and cross-border cultural transmission. This chapter focuses on just one dimension of globalization: international trade. The question is important because the political feasibility of maintaining open international markets for goods and services depends on the anticipated shares of specific groups as well as the aggregate benefits to the nation. The chapter reviews theory and empirical evidence on the likely consequences of expanded trade for inequality and poverty worldwide, within the United States, and within developing countries. It concludes with a discussion of the ability of policy makers in the United States and other countries to moderate pressures for a reversal of globalization trends.

## Defining, Measuring, and Evaluating Inequality

### Inequality Across What Population?

Most studies linking changes in inequality to globalization or increased trade focus on inequality within a single economy, that is, among households or, often, among workers employed in manufacturing industries. Some also look at relative earnings by race or gender, where inequality trends do not necessarily mirror those for the entire population. However, other researchers have emphasized the impact of globalization across nations or at the individual level for the world population. The focus is important because inequality across countries may be declining at the same time that inequality within individual countries is rising.

China and India offer clear examples of how answers to seemingly similar questions may be quite different. Both countries have recently experienced high rates of growth, in both cases resulting in a major reduction in the fraction of their population living in poverty. Growth has also translated to corresponding gains to an "average" individual in those countries as measured by gross domestic product (GDP) per capita, a figure that is often used in comparisons across countries. Moreover, rapid increases in GDP per capita have reduced the gap between these still-poor countries and the much richer nations of the Organisation of Economic Co-operation and Development (OECD); global inequality measured on this basis has thus fallen. But *within* most of the world's nations, whether rich or poor in terms of GDP per capita, inequality has risen. In the United States, inequality in U.S. incomes has been increasing since 1980, most recently due to increased wage dispersion at the very top of the distribution (Goldin & Katz, 2007; Piketty & Saez, 2003).

Taking individual *nations* as the unit of observation in an analysis gives equal weight to the world's largest countries and smallest mini-states. This may be appropriate for some purposes, as when evaluating the effectiveness of different types of political systems or economic policies in achieving sustained growth. However, it is less satisfactory as a way to evaluate how poor *people* (individuals and households) have been affected worldwide. India alone has a larger population than the 53 nations of Africa, and China's population is even larger. Together, China and India account for a major share of the world's poor but also of the dramatic recent reduction in the number of poor worldwide. Moreover, these reductions have occurred as the two economic giants have opened their economies to international markets. In part because of what has happened in India and China, the "world inequality level" among households as measured by a Gini coefficient[1] has trended downward since 1973 and by 2000 was nearly the same as in 1910 (Bhalla, 2002).

### Inequality of What?

Inequality may be measured in terms of a wide variety of economic outcomes. Within a country, the yardstick used most frequently is household (or family) income. A household unit often includes elderly persons or children with little or no earnings of their own. Their material well-being thus depends largely on earnings of others in the household. One problem with this approach is that household formation is itself sensitive to economic and demographic conditions. Maintaining a separate household may be a luxury that only the more affluent can afford. Moreover, the incentive for formation of new households depends on the age composition of the population as well as rates of marriage and divorce. Times-series evidence on trends in inequality measured at the household level must therefore be interpreted with care.

A second problem is how to evaluate the economic status of the household unit. Should the measure be monetary earnings? If so, should these be limited to earnings from current employment? Or should other types of income, such as returns on invested capital, business profits, and pension benefits, also be included? Should the measure be gross earnings or earnings net of taxes and cash transfers? In the United States and most other countries, net income is more equally distributed than gross income. What adjustment should be made for the "household production" of those family members whose work is performed in the home or for the market value of the services provided by durable assets (e.g., home, car) owned by the household? And what about the value of goods and services (e.g., food, housing, medical care) directly provided to the household by government units or other institutions?

These considerations suggest an alternative basis for measuring well-being: household consumption. But while consumption is more accurate than income as a means to evaluate the material well-being of a household, it is far less tractable from the point of view of statistical analysis. Data on incomes are largely drawn from statistics routinely collected by government units for other purposes, such as tax records, and coverage is often nearly universal. Household consumption data must be generated by surveys intended specifically for this purpose, with results based on statistical sampling of the population. Moreover, survey instruments and sampling techniques are subject to frequent revision, raising additional problems in evaluating trends over time.

To assess the impact of trade on inequality, researchers often use a much narrower definition: wage inequality, or even inequality of wages within the manufacturing sector and especially the relative earnings of skilled over unskilled workers, often defined as nonproduction relative to production workers. A major advantage of using manufacturing wages is that these can be matched to industry characteristics and exposure to trade. To the extent that trade does affect income or consumption inequality, the impact is most likely to come through changes in wages earned. Yet these changes can occur not only through changes in wage rates but also in employment and hours.

Trade liberalization typically causes some manufacturing sectors to shrink and others to grow. Over time, most industrialized nations have seen manufacturing jobs shrink as a share of total employment, with corresponding growth in the share of services employment. Accordingly, trends in the inequality of wages of manufacturing workers are of declining relevance as a measure of the overall redistributive consequences of expanded trade. An analysis focusing only on wages in manufacturing industries also misses income changes that occur when workers are displaced from manufacturing and subsequently employed in services, where average earnings are lower. For developing countries, manufacturing wages may be even more flawed as a base for evaluating changes in overall inequality. Rural families relying on self-employment or working for wages outside manufacturing are omitted, even though they are a disproportionate share of the population at the low end of the income distribution.

Regardless of the measure chosen, gaining an accurate account of the economic status of the poorest and the richest households is likely to pose additional challenges. For the poor, official statistics are likely to miss some or all income earned in the "informal sector" in developing countries and the "underground economy" in richer nations. Keeping employment off the books may be a means to avoid taxes or, in many cases more important, to maintain eligibility for income-tested government benefits; some unreported income is earned by individuals not legally eligible for employment, such as undocumented immigrants and underage workers. But such households may also be reluctant to provide a government employee with accurate consumption data. Surveyors may even find it too difficult or dangerous to locate such households—some at the very bottom of the income distribution are homeless, while others live in areas where crime is rampant. At the other end of the income distribution, tracking incomes of the rich can be limited by topcoding (which assigns an arbitrary maximum income value to preserve confidentiality of individual responses) and by complex legal instruments used by the rich to minimize taxes.

## Is Inequality Bad?

Most writers begin from the presumption that inequality is bad in itself, evidence of some underlying unfairness in the economic system. International organizations such as the World Bank often evaluate trends over time within a country in terms of their effect on the country's Gini coefficient or another measure of income inequality. In questioning inequality reduction as a social goal, Feldstein (1998) contrasts the conclusions regarding social welfare to be drawn from the Pareto criterion—that a change is good if it makes someone better off without making anyone else worse off—versus the Gini coefficient criterion—that a change is good if it lowers the Gini coefficient. When a change benefits those who are already best off without affecting others, the Pareto criterion judges it to constitute an improvement in social welfare. However, because it represents an increase in inequality, according to the Gini coefficient or similar measures, it is judged to worsen social welfare; given its distribution, the impact of a rise in the economy's total resources is then seen as having a *negative* impact from the social perspective.

Of course, few economic changes create benefits for some but losses for none. Most economic gains, no matter how significant in total, benefit some while hurting others. Certainly this is true of expanded trade as well as technological advance. In such cases, the likely redistributive consequences are indeed germane for public policy. However, it is important to distinguish situations in which inequality rises due mainly to increases at the very top of the income distribution (as in the United States and some other countries in the late 1990s and early 2000s) from those in which the rise in inequality is due at least partly to an absolute decline in the economic status of those already at the lower end in the income distribution.

Moreover, some degree of inequality is intrinsic to the efficient functioning of the market system. In particular, wage premiums earned by those with superior education, training, and job experience provide incentives to invest in those forms of "human capital." Likewise, profits earned by successful entrepreneurs provide the incentive to engage in new and usually risky ventures. Differences across individuals in rate of time preference imply that any measure of their economic well-being at a point in time will also differ even if each individual begins with an identical endowment and each maximizes expected lifetime utility. Because higher income households are

likely to save a larger share of total income, the distribution of income and wealth may also affect the savings rate in an economy and thus its rate of capital accumulation and growth.

Also discussed in the literature on economic inequality are the related issues of individuals' and households' mobility within the income distribution and inequality of outcomes versus inequality of opportunity. Society's concern regarding those at the low end of the income distribution is greater if their status is permanent rather than transitory. The evidence on these considerations provides some comfort, at least for the United States. A recent U.S. Treasury study (U.S. Department of the Treasury, 2007) finds significant mobility over time between income groups at both the top and the bottom of the U.S. income distribution. However, family income and wealth remain major determinants of children's educational attainments—and thus of their future earnings.

## Inequality Versus Poverty

Inequality refers to the properties of the entire distribution of income or another measure of economic well-being over all individuals or households in a country or other economic unit. In contrast, poverty refers to the status of a bottom group in the distribution. This group may be defined in terms of an absolute standard, such as $1 a day (extreme poverty) and $2 a day (moderate poverty) yardsticks used by the World Bank to measure progress in eradicating poverty in the poorest nations. The World Bank's goal in using these absolute measures is to treat two people with the same purchasing power equally (and judge them either to be poor or not poor) even if they live in different countries (Chen & Ravallion, 2008).

In contrast, the poverty lines countries use to evaluate their own progress in poverty eradication over time vary, with richer countries typically adopting a higher standard. The United Nations (2008) *Human Development Report 2007–8* uses explicitly different standards for developing countries and high-income OECD countries. Their broad "human poverty index" looks at (a) likeliness of surviving to age 40 (60) in developing (OECD) countries, (b) percentage of adults who are illiterate, and (c) material living standard. The last is measured for developing countries by percentage without access to safe water and percentage of children underweight for their age, but in the OECD group by the percentage below 50% of median household disposable income in that country—which is also the standard used by the OECD and the European Union in defining poverty. Thus, especially for wealthier countries, the line between poverty and inequality is blurred. A rise in incomes of others can result in a higher poverty line and thus an observed increase in the incidence of "poverty," even with no decline in the material well-being of those at the bottom of the income distribution. Measured trends in poverty rates may also be affected by the use of the same price deflators for all income levels, even though typical consumption weights differ by income.

For very poor countries with a significant part of the population already living close to the edge of subsistence, the redistributive consequences of increased trade (and of other types of economic changes resulting from policy decisions) are of fundamental significance. However, this is not because they may increase inequality but because they may increase or exacerbate poverty. Critics of efforts by the World Trade Organization (WTO), World Bank, and International Monetary Fund to promote trade liberalization in developing countries often argue that the poor do not benefit from any resulting increase in growth. In theory, a country could enjoy sustained rapid growth without any benefit to its poorest households if income disparities grew significantly—in other words, if the rich got richer while the incomes of the poor stagnated or declined. However, evidence suggests precisely the opposite. On average, World Bank researchers find that openness to foreign trade benefits the poor to the same extent that it improves overall economic performance (Dollar & Kraay, 2002). Even Rodrik (2000), skeptical regarding trade liberalization as a panacea for developing countries, acknowledges strong evidence that the poorest people usually benefit whenever a country gains overall: The number of people living in poverty has declined in every developing country that has sustained rapid growth over the past few decades. For the developing world as a whole, increased integration with global markets has been accompanied by a record decline in the percentage of poor people (Bhalla, 2002).

But increased trade is not invariably accompanied by a higher rate of economic growth. While the growth "success stories" among developing countries are all associated with increased trade (and in most cases with increased foreign direct investment), the empirical evidence that trade liberalization promotes growth is mixed. A path-breaking but controversial study by Sachs and Warner (1995) uses a 0–1 variable to characterize each country as either "closed" or "open" in a given year. This approach is sufficient to link openness to faster growth for the 1970s and 1980s but is less successful in explaining events of the 1990s. A follow-up study by Wacziarg and Welch (2008), based on improved openness measures and an expanded data set covering 1950–1998, finds an average increase in subsequent growth rates of about 1.5 percentage points per year for countries that liberalized trade, relative to their own preliberalization period. But this impressive positive impact is an average and masks considerable variation in the experience of individual nations. Because significant trade liberalization is usually just one of many policy changes made as part of a broader program, the heterogeneity in outcomes for growth needs to be evaluated in terms of the accompanying circumstances and policies; Wacziarg and Welch identify political instability, contractionary macroeconomic policy, and policies intended to shield domestic producers from adjustment to increased

import competition as confounding factors in the low- or no-growth cases.

## Effects of Opening to Trade on Inequality: Theory

Ricardo's nineteenth-century exposition of comparative advantage remains the basis for economists' prediction that a country must benefit from opening to trade. Despite innumerable theoretical refinements to standard trade models, today most economists remain confident that, at least as a practical matter, trade liberalization is almost always beneficial to a country as a whole. A more contentious issue is the domestic distributive consequences of trade liberalization. In particular, how are workers (as opposed to capital owners) affected? And more recently, how are unskilled (as opposed to better educated workers) affected?

### Partial-Equilibrium (Industry-Level) Analysis

Looking only at a single product without considering links to other parts of the economy, trade liberalization reduces the domestic price of the product. A simple supply-demand analysis then predicts that domestic demand for the product will rise, domestic supply will fall, and imports will increase to fill the resulting gap. Moving back along the domestic industry's supply curve implies a reduction in total production spending, which translates into some combination of reduced employment of productive factors, lower payments to those factors, and lower profits to firms in the industry. The immediate "losers" from trade liberalization are productive factors closely tied to the import-competing industry. The immediate gainers are domestic consumers of the product. The negative effect on *real* incomes of workers employed in the industry depends on the relative size of any reduction in money earnings and the gains from an increase in the purchasing power of those earnings due to the lower price of the product. For workers employed elsewhere, there is a clear gain—their (unchanged) earnings can now buy more of the product. The effect on workers overall is therefore ambiguous—it cannot be determined without additional information.

### General-Equilibrium (Economy-Wide) Analysis

Partial-equilibrium analysis provided the foundation for economists' traditional view of the effects of trade on income distribution until the work of Stolper and Samuelson (1941). By adopting a general-equilibrium framework based on the two-good, two-factor Heckscher-Ohlin (H-O) model of international trade, Stolper and Samuelson were able to obtain the unambiguous prediction known as the Stolper-Samuelson (S-S) theorem: Trade liberalization benefits a country's abundant factor and hurts its scarce factor. Their proof, which centers on varying factor demand and fixed factor supply, is intuitively appealing. The change in relative prices brought about by trade liberalization causes domestic production of the import-competing good to fall and domestic production of the export good to rise. This implies a fall in relative demand for the scarce factor, which according to the H-O theorem is used intensively in production of the import-competing good, and thus a fall in its earnings. For the capital-abundant United States, trade liberalization would unambiguously benefit capital owners and hurt workers, thus increasing income inequality.

Stolper and Samuelson's (1941) proof depends on the assumption that an economy's factor supplies are fixed, but an alternative demonstration by Jones (1965) does not require fixed factor supplies. Jones derives the S-S theorem directly from the equilibrium condition that, under perfect competition and constant returns to scale, unit cost must equal price for each good produced. Moreover, Jones's approach yields a key insight that applies more broadly: When the price of a good falls (for any reason), its average total cost of production must also fall to restore equilibrium. Under the simplifying assumptions of the H-O model, the redistributive impact can be pinned down precisely, as in Stolper and Samuelson. But even in a completely general setting, Jones's dual formulation in terms of factor prices and unit input coefficients (the cost-minimizing amounts of each factor used to produce one unit of the good) remains a useful tool for enumerating potential impacts: lower returns to at least one of the factor inputs employed, lower unit factor-input requirements, or some combination of the two. In fact, improved productivity (i.e., lower input requirements) is one of the "dynamic" benefits often claimed for trade liberalization. If labor can be made more productive as a result of increased import competition, the need for a drop in wages to restore equilibrium may thereby be reduced or even eliminated. Melitz (2003) models trade-related productivity gains achieved through sorting among firms that are heterogeneous with regard to labor productivity. Increased competition allows the most productive firms to expand and pay higher wages, forcing less productive firms to contract or exit.

However, the S-S framework provides an answer to a somewhat different question than the one policy analysts usually ask. In addition to the simplifying assumption of just two traded goods and two productive factors, the H-O model depicts a *long-run* equilibrium: long enough for factors to move freely between sectors, thereby maintaining full employment of both factors and equalizing the earnings of each factor across sectors. Yet for policy makers, the paramount concern is usually the *immediate* and *short-term* impact of trade liberalization—what happens as adjustment to liberalization proceeds, rather than after it has been completed. The same qualification applies to the models considered below of the redistributive consequences of offshoring.

## Short-Run (Specific-Factors) Analysis

While the S-S theorem applies only when productive factors are freely mobile between sectors and a given factor's earnings equalized across sectors, a variant model with some factors specific to one sector may be more relevant in explaining the short-run redistributive impact of trade liberalization. As separately developed by Mayer (1974) and Mussa (1974), the specific-factors model can be viewed as a short-run version of H-O, in which one factor is freely mobile between sectors and two others are either immobile or sector specific. This model predicts that the short-run effect of trade liberalization is to benefit factors tied to the export industry and hurt factors tied to the import-competing industry. The latter result conforms to the popular intuition that, for example, both steel mill owners and steel workers will be hurt by increased competition from imported steel. More surprising is that the predicted change in the real earnings of the mobile factor is ambiguous, contradicting the widespread belief that ability to move or adapt is key to benefiting from trade liberalization.

Both the S-S model and the specific-factors model predict likely gainers and losers from trade liberalization, an important political-economy question. Which model performs better in explaining observed behavior in the political sphere? Because the S-S theorem is based on perfect factor mobility within a country, its implications are best understood as long-term tendencies. Even assuming that factor owners seek to maximize the present discounted value of their lifetime earnings, the more immediate impact, which is better captured by the specific-factors model, is likely to dominate. Magee (1994) summarizes empirical literature evaluating the Stolper-Samuelson and specific-factors models as predictors of political behavior. Consistent with Stolper-Samuelson, Scheve and Slaughter (2001) find evidence that respondents with less education are more likely to favor protection. However, they also find stronger support for protection from those who own homes (i.e., immobile capital) in import-impacted areas.

## Skill-Biased Technical Change and Offshoring

Two further theoretical developments have influenced the most recent literature on the redistributive impact of trade liberalization; both have been stimulated in part by the inability of the highly simplified S-S framework to explain recent trends in inequality. Contrary to the S-S prediction of a lower ratio of skilled to unskilled labor use that would be expected to accompany the higher relative earnings of skilled labor, the use of skilled relative to unskilled labor actually has been rising over time in the United States. Moreover, the same is true in other industrialized countries and also in many developing countries.

### *Skill-Biased Technical Change*

A possible explanation for the divergence of skilled and unskilled wages, and the one many labor economists favor, is that technological progress is biased toward the use of skilled labor. Skill-biased technical change could moderate or even reverse producers' tendency to economize on more costly skilled labor as the skill premium rises. As *The Economist* (January 18, 2007) puts the question as to whether trade or skill-biased technical change is responsible, "Should you blame China or your computer?" If the rate of skill-biased technical change exceeds the rate of growth of skilled relative to unskilled labor, the premium paid to skilled labor would be expected to rise over time. However, technical change is itself endogenous, and at least some critics of skill-biased technical change as an explanation of increased inequality argue that increased exposure to imports from low-wage countries may stimulate exactly this type of technological change.

One implication of Jones's (1965) formulation in terms of factor prices discussed above is that an improvement in technology in a particular sector has a redistributive effect similar to that of a rise in the good's domestic relative price. In either case, there is "room" for higher costs and thus higher factor rewards. To restore the required equality of cost and price (this is the zero-profit condition for equilibrium) across industries, the reward to the factor used intensively in the industry with technological progress must rise, while the rewards of the other(s) must fall. Holding output prices fixed (the small-country assumption), this is true regardless of bias in the technological progress. Thus, "factor bias doesn't matter; sector bias does" (Leamer, 1998). But although concentration of productivity improvements in industries that use skilled labor intensively could account for the observed increase in the skill premium, without a skill bias it does not explain why the relative use of skilled labor has been rising.

### *Offshoring*

A second recent development in trade theory reflects observed changes in the international location of production. Increasingly, trade facilitates a geographical dispersion not only of the production of finished goods on the basis of comparative advantage but also of the individual steps in a production process—what Grossman and Rossi-Hansberg (2006, 2008) describe as "trade in tasks." Part of the rapid growth in trade flows relative to GDP is attributable to a finer international division of labor, in which intermediate products may cross national boundaries multiple times before the finished product reaches its market. This phenomenon is often called outsourcing. However, *offshoring* is a more accurate term for the location abroad of particular steps in the production process, whether this means moving a step to a firm's own foreign subsidiary or contracting with an unrelated foreign firm. (In the industrial organization literature, outsourcing refers to a situation in which a firm uses intermediate goods or services provided by other firms, either domestic or foreign, rather than carrying out a particular production step on its own.)

The redistributive consequences of offshoring, and indeed of any trade involving intermediate as well as final

goods, are more complex than when only final goods are traded. Consider first a situation in which intermediate as well as final goods are traded internationally. Liberalization of import restrictions on an intermediate good (say steel) has a negative impact on the factors used in its domestic production but benefits domestic producers of import-competing final goods using the intermediate in production (say autos); the *effective-protection rate* on domestic production of the final goods rises. In contrast to the nominal tariff on an import, the effective-protection rate measures the net advantage (or disadvantage) to a particular production process from tariffs on both the output of the process and the inputs to the process. A country's broad trade liberalization program may thus *increase* the amount of protection afforded to a particular sector (autos) if the tariffs on the inputs (steel) used in the sector are reduced by a larger percentage than the tariff on the output. In such a case, even though trade is liberalized (and more inputs are sourced from abroad), the predicted redistributive effects for productive factors used in the auto industry are those associated with an *increase* in protection on the industry's output, including an increase in output and employment. But it is also the result that would follow if the industry experienced an improvement in productivity, as described above.

Additional possibilities arise when tasks as well as purchased inputs are traded. Consider a wide range of productive activities ranked in terms of their skill intensity. If all tasks are equally amenable to being performed abroad, a reduction in the cost of offshoring (e.g., due to improvements in international communications) should increase the extent to which cheaper foreign labor is substituted for domestic labor. Benefits from offshoring may also increase when international capital flows raise the relative productivity of foreign labor. For a country such as the United States, where unskilled labor is relatively scarce, offshoring would relocate the least skill-intensive tasks still performed by U.S. workers to a country such as Mexico, where unskilled labor is more abundant and the relative wage of unskilled labor accordingly lower. As a result, the range of activities carried out in the United States becomes more skill intensive. Yet the same is true in Mexico, because the newly offshored tasks are more skill intensive than those previously performed there. Offshoring thus raises the average skill intensity of production and the relative earnings of skilled workers in *both* countries (Feenstra, 2010; Feenstra & Hanson, 1996).

By incorporating offshoring in a Heckscher-Ohlin model, Grossman and Rossi-Hansberg (2006, 2008) distinguish three effects after long-run equilibrium is restored (i.e., after production patterns and price adjust to reestablish equality of unit costs and prices). The first is a productivity effect: Offshoring some tasks increases the productivity of the domestic industry, an effect that tends to benefit the factor used intensively in that industry. As Grossman and Rossi-Hansberg (2006) observe, this first effect is equivalent to technological progress that augments the productivity of the same type of labor at home. Thus, if offshoring occurs mainly in industries that use unskilled labor intensively (for the United States, import-competing industries), the productivity effect tends to raise the absolute and relative returns to unskilled labor. However, offshoring has two further effects that work in the opposite direction. In response to the offshoring industry's improved productivity, its production will tend to expand. For a small country, by definition, the resulting impact on international prices can be neglected. But for a large country such as the United States, the expansion will depress the relative price of the industry's output. If the expansion is in the offshoring country's import-competing industry, this means a term of trade improvement, which is beneficial for the country overall but reduces the real return to the intensively used factor via the Stolper-Samuelson effect. Finally, offshoring acts like an expansion of the supply of the relevant factor, again tending to depress its earnings.

These analyses of offshoring show that the effects depend critically on which types of tasks can be offshored (less vs. more skilled) as well as which sectors are most affected (import competing vs. exporting). Early offshoring by U.S. firms affected mainly less-skilled jobs and import-competing industries, as modeled by Feenstra and Hanson (1996). However, recent offshoring also affects medium-skilled jobs such as customer service and transcription and high-skilled jobs such as computer programming and accounting. Jobs most likely to be offshorable are ones where desired performance can be fully specified in advance—what have come to be called "routine" jobs. But even among routine jobs, some require physical proximity while others can be performed remotely. In contrast to labor-intensive manufacturing, where offshoring of routine unskilled jobs is well established, many of the service-sector jobs held by unskilled workers—such as lawn care, janitorial, and food service—are not susceptible to offshoring. Because much of skilled work is carried out using computers, with results easily transmitted electronically, many of the more routine types of skilled tasks—such as legal, accounting, and computer services—have begun to be offshored, a trend that will likely accelerate as international electronic communication continues to improve and the supply of suitably skilled workers in emerging nations continues to grow. Skilled jobs that require ongoing interaction and consultation in the workplace, including many types of managerial jobs, are less likely to be offshored.

## Empirical Evidence on the Links Between Inequality and Trade

Theoretical analyses necessarily abstract from most of the complexities of the marketplace; alternative plausible simplifications offer conflicting predictions on the division of the productivity gains associated with increased trade. Yet all trade models underscore that gains may not be shared equally, and some models indicate the possibility of losses in absolute as well as relative terms. Newer trade models

that incorporate offshoring also suggest that the distribution of gains (and perhaps losses) may be difficult or impossible to anticipate in advance, and these may change over time as the array of tasks for which offshoring is economically advantageous continues to expand.

The increase in inequality documented in both developed and developing countries in recent decades has stimulated an explosion of empirical literature attempting to evaluate the relative importance of various contributing factors. The role of increased trade is particularly relevant from a public-policy perspective because many see trade as an influence readily controlled by the nation's economic policies. Accordingly, most researchers attempt to separate increased trade from other recognized factors, such as declining unionization, increased immigration, a declining real minimum wage, and especially skill-biased technical change. However, these other factors may themselves be influenced by increased trade.

## U.S. Studies

Gordon and Dew-Becker (2008) review a vast body of analysis and evidence on rising U.S. inequality. Labor's overall share in U.S. national income tends to fall during cyclical upturns and rise during cyclical downturns; once data are adjusted for movements associated with the business cycle, Gordon and Dew-Becker find no significant change over the past two decades. However, the recent inequality debate is no longer about the shares of labor and capital—Piketty and Saez (2003) conclude that "the working rich have replaced rentiers at the top of the income distribution"—but about the division of labor's share among those at the top and those lower in the income distribution. In fact, much of the observed increase in inequality arises from changes at the very top 10%, 1%, and even 0.01% of the population. For trade specifically, Gordon and Dew-Becker find evidence that trade with low-wage countries does affect earnings of U.S. workers adversely, but because they report results from a range of studies, they do not attempt to assign a relative ranking to trade among all potential influences.

Several recent studies by international economists focus more explicitly on the role of trade. Lawrence's (2008) analysis centers on the gap for the period 1981–2006 between growth in the wages of blue-collar workers and the overall growth in labor productivity (output per hour of labor input for all worker categories) in the business sector. Lawrence's goal is to determine how much higher blue-collar wages would have been in the absence of increasing inequality. To begin with, he finds that 60% of the gap between the growth rates of blue-collar earnings and overall labor productivity reflects two technical issues: omission of benefits from wage data and use of different price deflators to translate nominal wages and output into real values. He attributes a further 10% of the gap to the rising relative skills and education of non-blue-collar workers. For the 30% remaining, Lawrence sees technical change as the major explanation, with trade accounting for only about a fifth. Lawrence also points out that the effects of trade on inequality have actually been falling over time, as the U.S. economy moves away from producing the goods that compete most directly with low-cost imports. Expanded imports of products no longer produced at home benefit consumers through lower prices. Moreover, Broda and Romalis (2009) conclude that poorer families may benefit disproportionately from expanded U.S. trade because the expansion has been concentrated in the types of goods that are more important in their total spending.

The trade and inequality literature has focused mainly on wage effects, but Lawrence (2008) devotes a chapter to the role of trade's contribution to worker displacement. Although noting that trade is less important than other factors that cause U.S. workers to lose their jobs, he cites Kletzer's (2001) finding that when workers are displaced, older, less-skilled workers with more years of tenure are likely to experience the largest declines in reemployment earnings. Of workers displaced between 1984 and 2000, only one third of workers displaced from import-competing industries found new jobs in manufacturing, and the largest drop in average earnings after reemployment was for displaced workers who moved into nonmanufacturing industries. Moreover, in many cases displaced workers, especially married women with limited geographic mobility, did not find new employment at all.

Ebenstein, Harrison, McMillan, and Phillips (2009) link industry-level data on offshoring by U.S. multinational firms, import penetration, and export shares with data on individual workers. An important aspect of their empirical analysis is that it focuses on specific occupations rather than sectors. The results confirm the importance of wage effects of trade and offshoring that operate across rather than within industries. Workers who remain in manufacturing are on average *favorably* affected by offshoring. Similar to Kletzer (2001), they find significant downward pressure on wages for those who leave manufacturing to take jobs in agriculture or services. Consistent with the theoretical results of Grossman and Rossi-Hansberg (2008), they find an ambiguous overall impact of offshore employment on domestic wages; offshoring to high-wage locations is positively associated with U.S. wages while offshoring to low-wage locations is not.

## Developing Country Studies

The Stolper-Samuelson theorem predicts that expanded trade in developing countries should benefit unskilled labor, the locally abundant factor, while hurting scarce skilled labor, thus reducing inequality. However, as more developing countries have become integrated into global markets and in many cases experienced rapid growth of per capita income, data on the distribution of economic gains from expanded trade show the opposite. In a survey of

empirical studies covering many developing countries over several decades, Goldberg and Pavcnik (2007) find substantial evidence of "a contemporaneous increase in globalization and inequality." However, some of these studies do not attempt to examine overall inequality but look only at wage earnings of those working in the formal sector or even only those working in the manufacturing sector.

In addition to considerations highlighted in U.S. studies, such as skill-biased technical change, Goldberg and Pavcnik (2007) point to constrained labor mobility and other factor–market distortions that can inhibit adjustment to changed economic incentives. Where the adjustment process is inhibited, both the increase in inequality but also the total gains are likely to be smaller. Moreover, trade liberalization in developing countries is typically part of a broader policy package, complicating the problem of identifying a causal relationship between increased trade and increased inequality.

Though focused on the relationship between globalization and poverty, most of the cross-country and single-country studies in Harrison (2006) also address the trade-inequality link. Harrison's introductory essay cautions that some of the poor, especially those employed in import-competing sectors, will indeed lose from trade liberalization. The country studies highlight the central role of complementary policies, such as those addressing deficits in education, infrastructure, and access to credit, in achieving overall gains. Especially important is the ability of workers to relocate from contracting sectors into expanding ones. Likewise, given the inevitable dislocations associated with adjustment to trade liberalization, "careful targeting" is necessary to protect the poor from negative consequences. But notwithstanding concerns about possible increases in poverty or inequality, especially during what may be a protracted adjustment period, Harrison underscores the importance of improving the access of exporters in developing countries to the markets of the affluent developed countries, pointing to "a clear link between export activity and poverty reduction" documented in several of the country studies.

## Can Globalization Be Sustained?

Higher efficiency and faster growth are not ends in themselves. They simply increase the total resources potentially available to achieve society's preferred goals. Likewise, policies to facilitate globalization (i.e., to achieve greater openness and international integration) do not by themselves ensure progress toward social goals. To be sure, some of the recent backlash to globalization is simply the expression of private interests. As with any important advance in technology, the gains achieved through globalization are accompanied by powerful redistributive consequences associated with the restructuring of firms, industries, and entire economies. Moreover, it is usually easier to predict who will be the losers from trade liberalization than who will be the winners. Democratic systems give potential losers the power to hold change hostage, to insist on protection or compensation as the price of their assent.

For reasons of both fairness and political feasibility, maintaining the momentum of efficiency-promoting change requires a mechanism for ensuring that the gains are broadly shared. Rodrik (1997) documents a positive relationship between government spending and openness in the OECD countries. He interprets this pattern as indicating that greater exposure to external market forces requires a more active government role to cushion losers and thus ensure a socially and politically acceptable sharing of gains. Likewise, safeguard and antisubsidy provisions in trade agreements, consistent with the rules of the World Trade Organization but often criticized as protectionist loopholes, serve as economic shock absorbers that may be politically necessary if national governments are to liberalize access to their domestic markets.

To the extent that globalization entails redistribution among countries as well as within them, provision of social insurance only at the national level may be inadequate to stave off protectionist responses. The European Union's successful expansion of an integrated multinational market has required a mechanism for sharing benefits across as well as within national boundaries. If the gains from globalization are large enough, other nations may likewise be willing to cede authority to an international body in order to maintain them, but a more likely outcome is further liberalization along regional lines. As competing regional groups form, they may develop alternative approaches to balancing efficiency gains from integration with mechanisms for ensuring an acceptable division of benefits among members.

Much of the discussion of globalization's effects concerns the potential for a policy backlash that could reverse the recent trend toward greater integration. We learn from the experience of a century ago that despite its significant contribution to national economic performance, globalization is highly vulnerable to political factors. It is surely no coincidence that the Great Depression unleashed protectionist policy changes in the most important nations around the world. But while the same kinds of redistributive pressures are evident today, most national governments are now better equipped to maintain the viability of openness by ensuring a politically acceptable sharing of its economic benefits within nations and even within regions. Moreover, the rules of the World Trade Organization now help to restrain the protectionist responses of member nations, as has been seen during the global recession. The absence of serious breaches in members' commitment to WTO rules even in the face of the worst economic downturn the world has experienced since the Great Depression is a reason for optimism that another 1930s-style reversal of the trend toward globalization can be averted.

## Note

1. The Gini coefficient is a standard statistical tool often used to compare inequality across countries at a point in time or in a single country over time. It is based on a Lorenz curve, which shows cumulative share of total income (or expenditure or consumption) on the vertical axis and cumulative share of total population (or total households) on the horizontal axis. For a country in which economic resources are distributed equally, the Lorenz curve is a 45-degree line. The Gini coefficient is computed as twice the area between the 45-degree line and the actual Lorenz curve or, equivalently, the ratio of the area between the 45-degree line and the Lorenz curve and the area below the Lorenz curve. The Gini coefficient is therefore between 0 and 1, where 0 corresponds to perfect equality and 1 corresponds to perfect inequality (one person or household gets everything). A simpler measure often used to measure extremes of inequality is the ratio of the bottom decile (or quintile) to the top decile (or quintile). This measure provides no information regarding those in the middle of the distribution.

## References and Further Readings


Bhalla, S. S. (2002). *Imagine there's no country.* Washington, DC: Institute for International Economics.

Broda, C., & Romalis, J. (2009). *The welfare implications of price dispersion* (Working paper). Chicago: University of Chicago.

Chen, S., & Ravallion, M. (2008). *China is poorer than we thought, but no less successful in the fight against poverty* (World Bank Policy Research Working Paper WPS 4621). Washington, DC: World Bank.

Dollar, D., & Kraay, A. (2002). Growth is good for the poor. *Journal of Economic Growth, 7,* 195–225.

Ebenstein, A., Harrison, H., McMillan, M., & Phillips, S. (2009). *Estimating the impact of trade and offshoring on American workers using the Current Population Surveys* (NBER Working Paper 15107). Cambridge, MA: National Bureau of Economic Research.

Feenstra, R. C. (2010). *Offshoring in the global economy: Microeconomic structure and macroeconomic implications.* Cambridge: MIT Press.

Feenstra, R. C., & Hanson, G. (1996). Foreign investment, outsourcing and relative wages. In R. C. Feenstra, G. M. Grossman, & D. A. Irwin (Eds.), *Political economy of trade policy: Essays in honor of Jagdish Bhagwati* (pp. 89–127). Cambridge: MIT Press.

Feldstein, M. (1998). Is income inequality really a problem? In Federal Reserve Board of Kansas City (Ed.), *Income inequality: Issues and policy options* (pp. 357–367). Kansas City, MO: Federal Reserve Bank of Kansas City.

Goldberg, P. K., & Pavcnik, N. (2007). Distributional effects of globalization in developing countries. *Journal of Economic Literature, 45,* 39–82.

Goldin, C., & Katz, L. F. (2007). Long-run changes in the U.S. wage structure: Narrowing, widening, polarizing. *Brookings Papers on Economic Activity, 2,* 135–165.

Gordon, R. J., & Dew-Becker, I. (2008). *Controversies about the rise of American inequality: A survey* (NBER Working Paper 13982). Cambridge, MA: National Bureau of Economic Research.

Grossman, G. M., & Rossi-Hansberg, E. (2006). The rise of offshoring: It's not wine for cloth anymore. In Federal Reserve Board of Kansas City (Ed.), *The new economic geography: Effects and policy implications* (pp. 59–102). Kansas City, MO: Federal Reserve Bank of Kansas City.

Grossman, G. M., & Rossi-Hansberg, E. (2008). Trading tasks: A simple theory of offshoring. *American Economic Review, 98,* 1978–1997.

Harrison, A. (Ed.). (2006). *Globalization and poverty.* Chicago: University of Chicago Press.

Jones, R. W. (1965). The structure of simple general equilibrium models. *Journal of Political Economy, 73,* 557–572.

Kletzer, L. G. (2001). *Job loss from imports: Measuring the costs.* Washington, DC: Institute for International Economics.

Lawrence, R. Z. (2008). *Blue-collar blues: Is trade to blame for rising US income inequality?* Washington, DC: Peterson Institute for International Economics.

Leamer, E. E. (1998). In search of Stolper-Samuelson effects on U.S. wages. In S. M. Collins (Ed.), *Imports, exports, and the American worker.* Washington, DC: Brookings Institution Press.

Magee, S. P. (1994). Endogenous protection and real wages. In A. Deardorff & R. Stern (Eds.), *The Stolper-Samuelson theorem: A golden jubilee* (pp. 279–288). Ann Arbor: University of Michigan Press.

Mayer, W. (1974). Short-run and long-run equilibrium for a small open economy. *Journal of Political Economy, 82,* 955–967.

Melitz, M. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica, 71,* 1695–1725.

Mussa, M. L. (1974). Tariffs and the distribution of income: The importance of factor specificity, substitutability, and intensity in the short and long run. *Journal of Political Economy, 82,* 1191–1203.

Piketty, T., & Saez, E. (2003). Income inequality in the United States, 1913–1998. *Quarterly Journal of Economics, 118,* 1–39.

Rodrik, D. (1997). *Has globalization gone too far?* Washington, DC: Institute for International Economics.

Rodrik, D. (2000). Growth versus poverty reduction: A hollow debate. *Finance and Development, 37*(4), 9–10.

Sachs, J. D., & Warner, A. (1995). Economic reform and the process of global integration. *Brookings Papers on Economic Activity, 1,* 1–118.

Scheve, K. F., & Slaughter, M. J. (2001). *Globalization and the perceptions of American workers.* Washington, DC: Institute for International Economics.

Stolper, W., & Samuelson, P. A. (1941). Protection and real wages. *Review of Economic Studies, 9,* 58–73.

United Nations. (2008). *Human development report 2007–8.* New York: United Nations Development Programme.

U.S. Department of the Treasury. (2007). *Income mobility in the U.S. from 1996 to 2005.* Available from http://ustreas.gov/offices/tax-policy/library/incomemobilitystudy03-08revise.pdf

Wacziarg, R., & Welch, K. H. (2008). Trade liberalization and growth: New evidence. *World Bank Economic Review, 22,* 187–231.

# 49

# The Economics of Fair Trade

John D. Messier

*University of Maine at Farmington*

After oil, coffee is the most traded commodity in the world. Like all commodities, coffee prices have historically been quite volatile. Coffee, unlike other internationally important commodities, is frequently produced by small family-run operations. As such, volatile and unpredictable coffee prices leave coffee producers in a vulnerable and precarious position. Furthermore, coffee is an important source of export earnings and foreign exchange for many countries, particularly developing countries. This is due to the geography of coffee. Coffee is a tree crop that thrives when grown at an altitude along the equator, situating it principally in developing countries. Fluctuating coffee prices create income uncertainty for households and nations that rely on coffee earnings.

There is a long history of intervention in commodity markets to control prices. The Organization of the Petroleum Exporting Countries (OPEC) is a well-known example of intervention to manipulate prices. Fair trade is an alternate trade arrangement that aims to both raise earnings and reduce income variability for coffee producers. Instead of manipulating supply as OPEC does, fair trade is a voluntary market-based response allowing consumers to express their preferences by choosing to pay a premium for goods produced in accordance to fair trade standards. As such, fair trade can be seen as a social response to market forces and needs to be understood as a response to poverty and globalization.

## Background

There are two types of coffee beans: Arabica and Robusta. Arabica beans are generally considered to be of higher quality and superior taste. Both types of coffee grow in tropical climates. Arabica is grown at higher altitudes, typically between 3,000 to 6,000 feet above sea level with temperatures between 59 and 75 degrees Fahrenheit year round, while Robusta grows from sea level to 3,000 feet. Robusta is more tolerant of heat, but both are very sensitive to rainfall and frost. Coffee requires roughly 60 to 120 inches of rainfall per year with a dry season during which the harvesting occurs (Wrigley, 1988). These conditions position prime coffee production along the equator in developing countries.

Coffee is a tree crop, and the trees take 3 to 5 years to mature before they bear viable fruit. Coffee itself is the seed contained within the fruit from the coffee tree. The fruit is harvested then pulped to extract the seeds, which are then dried for export as green beans. The lag between planting and coffee production is one of the factors that often leads to price instability. When prices are high, growers have an incentive to expand production. However, prices can remain high for a few years before supply can be increased, as farmers plant additional trees and wait from them to mature. Given the high level of poverty and limited income opportunities in most coffee-producing countries, price increases provide strong incentives to increase output. When the trees mature, supply rapidly increases, and all else constant, price falls. On the other hand, when prices are low, many farmers can no longer afford to grow coffee. They begin to neglect the trees or even destroy them as they move to a new crop. This leads to a cut in supply, and a new cycle begins.

Figure 49.1 shows the price in U.S. cents per pound of green coffee. The composite price is constructed from daily averages of green coffee from the three major global coffee markets: New York, Germany, and France. Prices range from a high of 314.96 cents per pound in 1977 to a low of 37.22 cents per pound in 1967.

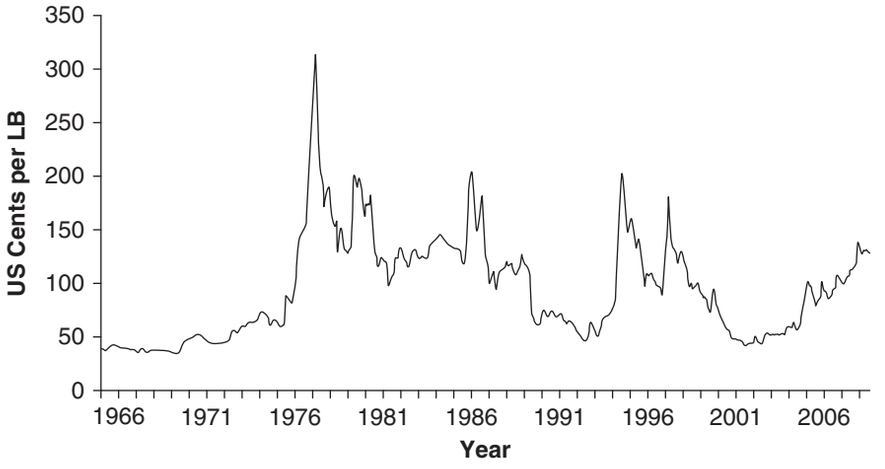


**Figure 49.1** Composite Indicator Price of Green Coffee

SOURCE: Author's calculation based on data from International Coffee Organization (http://www.ico.org).

Brazil, one of the world's largest producers of coffee, experienced a severe frost in 1975. The Instituto Brasileiro do Café (IBC) reported that 50% of the crop had been destroyed. The frost occurred after harvesting had begun, so the effect on prices was not felt until the following year. With a drastic reduction in supply, prices spiked upward in 1976 and 1977. The price increase encouraged an expansion of coffee planting, both to replace the damaged plants and to try to take advantage of higher prices. The first of this coffee began to enter the market in 1980, increasing supply and resulting in falling prices. Further boom and bust cycles were experienced in 1986, 1997, and 1999 due to droughts in Brazil and in 1994 when Brazil was hit by another frost.

These price cycles leave producers and nations that rely on coffee revenue in a precarious position. Low and unstable income limits households' ability to save, acquire assets, or invest in their production. With little or no savings or assets, households cannot borrow to smooth consumption in economic downturns. Lacking savings or access to credit, households are forced to cut consumption and may remove children from school as coffee prices fall. Low and unstable income also leaves individuals vulnerable to illness or accidents. Cuts to consumption lead to poor nutrition and inadequate health care, increasing the likelihood of illness, which can be devastating to the household. The effects of illness and injury are twofold. Caring for the ill or injured creates an additional burden due to medical costs. Also, the opportunity cost of lost earnings for the caregiver and the injured or ill can be profound.

The impact of price fluctuations is not restricted to households. Nations relying on coffee exports for government revenue and foreign exchange are similarly affected. Falling government revenues necessitate cuts in spending, particularly for social support programs such as access to health services, education, and nutritional aid. These cuts in government spending fall disproportionately on the poor and most vulnerable in society.

Weather is not the only source of instability in the coffee market. Global supply and demand conditions also affect coffee prices. Recessions in coffee-drinking countries trickle down to lower prices for farmers. These green coffee prices are set in commodity markets, such as the New York Board of Trade (NYBOT).

The NYBOT serves as the industry benchmark for green coffee beans and clears the market based on future contracts. The Inter-Continental Exchange (ICE) coffee futures market was established in 1882 to promote price stability in the coffee industry. According to the ICE, the coffee market has exhibited greater historic volatility than other commodity markets. This volatility leads to more hedging as traders look to avoid price risk. Coffee traders can reduce risk by buying contracts that fix the price for coffee to be delivered in the future. Traders can also speculate in the coffee market, increasing price volatility. Unlike traders, most coffee growers do not have access to insurance or other markets that would allow them to reduce risk.

Various attempts to reduce price and earning volatility for coffee growers have been instituted over time. Fair trade is the latest in a serious of pro-poor trade regimes. Fair trade is a partnership between coffee consumers in relatively wealthy nations and coffee growers in poor nations. It is designed to promote greater equity in international trade and sustainable development while protecting workers rights. Fair trade operates as a certification process. International certifiers require producers to meet certain standards to qualify for the fair trade label. The fair trade model can be applied to a variety of goods; however, it is principally applied to edible commodities. This chapter will focus on coffee, the mainstay of the fair trade movement.

## Origins of Fair Trade

The rise of fair trade came out of the confluence of two events: changes in global market conditions and the rising

international social justice movement. Global markets were dramatically affected by the rise of structural adjustment policies leading to prolonged and often painful economic restructuring. In the same period, there was growing awareness of the immiseration of the working poor globally.

### *Structural Adjustment Policies*

Structural adjustment was a set of policies promoted principally by the World Bank and International Monetary Fund (IMF) designed to stabilize countries' balance of payments. Following the classical model of international trade, nations focused on areas of comparative advantage. This resulted in industrial countries focusing on industrial goods while poorer, less developed countries relied on agricultural and unprocessed primary goods. Trade in this manner should, according to standard economic models, lead to falling opportunity costs for imported goods and therefore rising standards of living and welfare for all nations.

This premise was called into question by Raul Prebisch, executive secretary of the United Nations Economic Commission for Latin America and the Caribbean (ECLAC) and later secretary-general of the United Nations Conference on Trade and Development (UNCTAD). Prebisch (1971) showed that the terms of trade, the quantity of foreign goods (imports) that were exchanged for a unit of domestic goods (exports), were falling. The implications are that foreign, manufactured goods were becoming more expensive. As each export from the developed world was exchanged for ever increasing amounts of primary goods from developing countries, trade was seen as exploitive—leading to higher standards of living in the industrial countries but not in the less developed countries.

Prebisch (1971) noted some of the causes of the declining terms of trade—in particular, falling income elasticity of demand for primary goods and growing protection of domestic markets in developed countries. Industrialized nations experienced rapid gains in productivity in the 1950s. The increase in output generated an expansion in international trade as markets were found for the additional output. This led to rising wages and income in the industrial countries. As incomes rose, demand for manufactured goods also increased, further expanding income in the industrialized countries.

Unfortunately, producers of primary goods in developing countries did not benefit from the rising income in industrial countries. As incomes were rising in industrial countries, the income elasticity of demand for primary goods was falling. Demand for output from developing countries was not keeping pace with demand for manufactured goods. Prices of manufactured goods were rising faster than primary goods, widening the income gap between First World and Third World nations and contributing to the declining terms of trade. At the same time, many industrialized countries bowed to domestic pressure and imposed trade restrictions on many agricultural and light manufactured goods. Producers in developing countries faced the twin threats of declining demand for their goods and reduced market access for exports.

Prebisch (1971) concluded that prices for goods from the industrialized nations would continue to rise relative to the less developed nations. Faced with declining terms of trade, economic growth and development through trade seemed impossible. For Prebisch, trade clearly did not make everyone better off. While trade would lead to material improvements for wealthy nations, the majority of the world would face immiseration, producing raw materials or lightly processed primary goods for the developed world, resulting in low productivity and pay for Third World nations for the foreseeable future.

For Prebisch and others, the road to higher standards of living required industrialization. Industrialization, however, would raise living standards in developing countries only if there was demand for output. Given the differences in capital and productivity, there was no way these infant industries could compete with multinational corporations dominating international trade. Successful industrialization required intervention.

The intervention used in many developing countries was import substitution industrialization (ISI). ISI was a growth strategy that focused on domestic industrialization. Realizing that domestic industry was incapable of competing with experienced and well-funded multinational corporations, many developing countries focused on producing for the domestic market with the hopes of eventually being able to compete internationally. To gain a competitive advantage in the domestic market, governments adopted restrictive policies on imports. Governments used a variety of tools, including tariffs, quotas, and exchange rates, to raise the price of selected imports. This created space for domestic production. Taking advantage of reduced competition and government support, new industries emerged to fill the demand for manufactured goods.

The initial effects of import substitution were promising. Many countries reported economic growth, rising employment, and reductions in poverty. New challenges, however, soon emerged. The very strategies adopted to create space for domestic producers reduced the efficiency of exporters. The shift into new industries increased domestic competition for inputs. This led to rising wages and resource prices, increasing costs for traditional exports in the agricultural sector and ultimately inflation. Also, governments often replaced private industry with state-owned enterprises, contributing to rising budget deficits.

To continue industrialization, countries needed imports of capital goods—machines and intermediary goods that could not be produced domestically. To pay for the capital and infrastructure needed to industrialize as well as fund rising government deficits, countries expanded the money supply and borrowed internationally.

International debt requires repayment. To repay, countries must run a trade surplus to generate sufficient foreign exchange for debt service. That is, they need to export more than they import to have foreign exchange left over

for repayment. ISI policies, however, lead to the exact opposite situation. Rising costs of production and overvalued exchange rates made most nations' exports uncompetitive internationally. When global economic conditions slowed following the oil crises, demand for exports fell even further. At a time when developing nations needed massive amounts of foreign exchange, their uncompetitive exports generated little. The end result of the ISI experiment was many heavily indebted countries facing a balance-of-payment crisis.

Nations facing shortages of foreign exchange have few options. With a growing balance-of-payment crisis, nations either had to borrow more foreign exchange for debt service or default on their loans. Defaulting was not an attractive option as it affects a nation's ability to borrow in the future. Furthermore, there was tremendous international pressure from lending nations to ensure repayment. Borrowing was also problematic. The international banking community was unwilling to invest in more potential bad debt.

This void was filled by the IMF. The IMF was created in 1944 in Bretton Woods, New Hampshire, and one of its primary purposes is to provide temporary financial assistance to countries in need of balance-of-payment adjustment (www.imf.org). Assistance from the IMF was not without conditions. Borrowing countries were in the current predicament because they could not generate sufficient foreign exchange to repay international debt. IMF conditionality focused on restructuring the economy to generate more foreign exchange. The goal was to make exports more attractive (think cheaper) and imports more expensive. The Structural Adjustment Policy (SAP) era had begun.

Structural adjustment generally began with exchange rate rationalization. The struggling nations typically had fixed and overvalued exchange rates. The overvaluation was a result of inflation as well as other policy choices. With inflation, the average price of all goods, including exports, rises. As prices rise, the demand for a nation's exports falls, resulting in a decline in the demand for their currency. At the same time, imports become relatively cheaper than domestic goods, leading to an increase in demand for imports and foreign exchange. The end result for the struggling nation is an increase in the supply of its currency in pursuit of imports and a fall in demand of its currency in response to declining demand for exports—the price of the currency should fall to return to equilibrium. With a fixed exchange rate, however, this adjustment does not take place. The IMF's approach to overvalued exchange rates was twofold. First, nations needed to devalue their currency to more closely reflect market conditions. Second, they needed to address the source of overvaluation—inflation.

Much of the inflation of ISI nations was due to their prior policies—subsidizing industrialization, building infrastructure, and operating state-owned enterprises. From a macroeconomic perspective, government spending was generating demand push inflation. Reducing spending required cutting government budgets and privatizing state-owned enterprises—a policy-induced recession. The immediate effect was increased unemployment and poverty for the global majority, setting the stage for the fair trade movement.

### *Growth of Social Justice Movement*

The second component leading to the rise of fair trade was the collapse of the International Coffee Agreement (ICA). Commodity markets have historically gone through boom and bust periods. Oversupply leads to price collapse, recessions lead to reduced demand and price collapse, and favorable conditions such as good weather and appropriate rainfall lead to a surge in output and price collapse. Variability in both supply and demand leave commodity producers in general and coffee producers in particular in a precarious position. In an effort to dampen the boom and bust cycles experienced by coffee producers, the ICA was ratified in 1962 (www.ico.org).

The ICA was negotiated under guidance of the United Nations. The agreement reached in 1962 established a series of quotas among coffee-producing nations to stabilize prices. The excess supply was removed from the marketplace and destroyed, primarily by national coffee boards. The ICA established a threshold target price for coffee. When demand for coffee was sufficiently high and the price increases above the threshold, the quotas were relaxed and producer earnings increased.

In 1986, Brazil experienced a prolonged drought that reduced Brazil's output by 1.14 billion kilograms (U.S. Department of Agriculture [USDA], 1999). This massive contraction led to a price increase and the suspension of the quota by the ICA. Once prices returned to historical levels, it proved difficult to reinstate the quotas. Nations did not want to reduce their coffee exports and earnings, requiring prolonged negotiations to reintroduce the ICA. While the quotas were reinstated in the 1987 agreement, they were short-lived. By 1989, it became clear that price control though a quota system was unmanageable.

Part of the difficulty lay in the reliance on national coffee boards for verification and enforcement of the quota system. Most national coffee boards were either vastly underfunded or completely eliminated during the mandated structural adjustment cuts to government spending. With insufficient oversight and outright corruption, the flow of coffee was virtually unmanaged. Recognizing this, in 1989, the ICA shifted its focus away from price control toward market expansion and infrastructure development (www.ico.org/history.asp). The International Coffee Organization continues to promote coffee consumption and to provide practical support to the coffee community.

A second factor contributing to the collapse of the ICA was the fall of communism. The ICA was seen as a tool to promote Western ideology. The quota system as well as other forms of funding and aid led to higher incomes for

rural farmers. Higher and more stable income was hoped to reduce the incentives for farmers to join radical movements, particularly communist and Marxist groups. With the end of the cold war, the West's concern with leftist movements was reduced, and support for ratifying the ICA crumbled.

The rise of structural adjustment and the collapse of the international coffee agreement had a significant impact on coffee producers. The implementation of structural adjustment required governments to substantially cut spending. This resulted in rising unemployment and a reduction in social spending on health care, education, and other social programs. Furthermore, the move toward free trade and open economies led to a collapse of the industrial sectors in many developing countries. Free trade pushed countries to produce where there existed a natural comparative advantage, resulting in an expansion of traditional exports, typified by the movement into coffee production by many of the world's poor. As coffee production began to expand, the collapse of the ICA removed price guarantees for coffee producers, resulting in falling incomes, subsistence farming, and highly precarious living.

This void was filled by civil society, principally social justice groups. Fair trade markets find their roots in more than 50 years of alternative trade relationships. Long before certification existed, churches, disaster relief organizations, and solidarity groups had formed more direct trade relationships with refugees and marginalized groups. The goal of these organizations was to get more income into the hands of poor producers around the world. They worked to increase market access for Third World producers by connecting directly with the producers and creating new distribution chains through religious and solidarity networks. This direct connection with producers resulted in better prices and increased market access. The solidarity groups were also able to provided technical assistance and improve access to education and health care. The impact of the solidarity movement was limited by the number of solidarity groups and small markets they had access to, resulting in a low volume of fairly traded goods. The ability to promote development and improve quality of life for poor producers was limited.

To increase the impact, fair trade groups had to tap into existing markets. Solidarity and church groups were effective in matching socially minded consumers directly to producers but lacked the scale to have a significant impact. A solution was introduced by Solidaridad, a Dutch alternative trade organization. Solidaridad created a label, Max Havelaar, which guaranteed that the goods carrying the label met specific labor and environmental standards. The label was initially applied only to coffee. Labeling moved socially minded consumers out of church halls and basements and into supermarkets and other retail centers. The label provided a signal of the direct connection to producers that in the past was provided by word of mouth in church and solidarity groups.

The labeling initiative caught on. Max Havelaar was soon joined by the Fairtrade Foundation, Transfair, and Rattvisemarkt. Initially, the fair trade organizations operated independently, but in 1997, they joined together to form the Fairtrade Labeling Organizations International (FLO). FLO works to standardize fair trade, support fair trade initiatives, and certify producers.

### *Fair Trade Standards*

Fair trade labels ensure that certain standards are maintained by producers of fair trade products. Fair trade certifiers monitor producers to ensure that the standards are met. Remember that fair trade arose from the alternative trade movement. It is building a model that does not put profit maximization as the primary goal. Rather, the movement is focused on changing the relationship between producers and consumers. It hopes to raise awareness in consumers of the impact of their consumption on others, including the producers and the environment, while at the same time lifting poor producers out of poverty.

Below is a summary of fair trade principles (Clement & Defranceschi, 2006):

1. *Creating opportunities for economically disadvantaged producers.* Fair trade is a strategy for poverty alleviation and sustainable development. Its purpose is to create opportunities for producers who have been economically disadvantaged or marginalized by the conventional trading system.
2. *Transparency and accountability.* Fair trade involves transparent management and commercial relations to deal fairly and respectfully with trading partners.
3. *Capacity building.* Fair trade is a means to develop producers' independence. Fair trade relationships provide continuity, allowing producers and their marketing organizations to improve their management skills and gain access to new markets.
4. *Payment of a fair price.* A fair price in the regional or local context is one that has been agreed through dialogue and participation. It covers the costs of production and enables socially just and environmentally sound production. It promotes gender equity and equal pay for equal work by women and men. Fair trade ensures prompt payment and, when possible, provides credit for producers.
5. *Working conditions.* Fair trade promotes a safe and healthy working environment for producers. The participation of children (if any) must not adversely affect their well-being, education, or need for play and conforms to the UN Convention on the Rights of the Child as well as the law and norms in the local context.
6. *The environment.* Fair trade actively encourages better environmental practices and the application of responsible methods of production.

In the context of coffee, fair trade focuses on small-scale producers, generally family-owned farms. A driving goal of fair trade is to reduce poverty. Therefore, only

small-scale producers are eligible for certification, excluding plantation production schemes. Wholesale purchasers of fair trade coffee have minimum buying requirements; therefore, small producers must band together and form cooperatives. By pooling their output, cooperatives can attain economies of scale in production and sales. Fair trade coffee cooperatives also promote community involvement and can fund community-level projects such as the construction of roads or medical centers.

To use fair trade labeling, wholesalers must buy directly from approved grower organizations using purchasing agreements that extend beyond one harvest cycle. Wholesalers must also pay the FLO minimum price (U.S.$1.21 per pound for Arabica coffee), pay an additional $.15 per pound for coffee certified organic, and pay a social premium of $.05 per pound. To ease credit constraints, purchasers must also offer financing equal to 60% of the contract value. This arrangement creates long-term partnerships between producers and wholesalers and encourages growers to invest in infrastructure. It also allows for more direct access to markets and prevents predatory practices where buyers prebuy coffee harvests at low prices from desperate farmers.

## Theory

There is no single theoretical framework to model fair trade. Three strands of theory will be explored that can be used to better understand fair trade. First is a discussion of demand for fair trade coffee based on utility and consumer choice. This is followed with a discussion of the supply decision for fair trade coffee growers. Finally, fair trade is presented in the framework of market failures and global value chains in coffee production.

Standard economic modeling describes the problem of consumer choice as one of utility maximization. Rational agents will spend each dollar to provide the greatest amount of utility possible. The maximizing agent will be in equilibrium when the marginal utility divided by the price is equal across goods. In this light, the purchase of fair trade coffee would be difficult to justify as it is providing the same consumption experience at a higher price.

Consider two goods: good $x$ (say, coffee) and good $y$, a bundle of all other goods. For a given level of income, consumers aim to maximize their utility subject to their budget constraint, as shown in Figure 49.2. For given income and prices, consumers will buy the last bundle of goods affordable—where the indifference curve is just tangent to the budget line ($BL_1$). This is shown at point A in Figure 49.2. Now consider the move to fair trade coffee. Because fair trade coffee is more expensive than conventional coffee, the budget line rotates to $BL_2$—the same income now buys less. Indifference curve 1 (IC1) is no longer affordable, and the consumer moves to a lower level of total utility (IC2), a seemingly irrational choice.

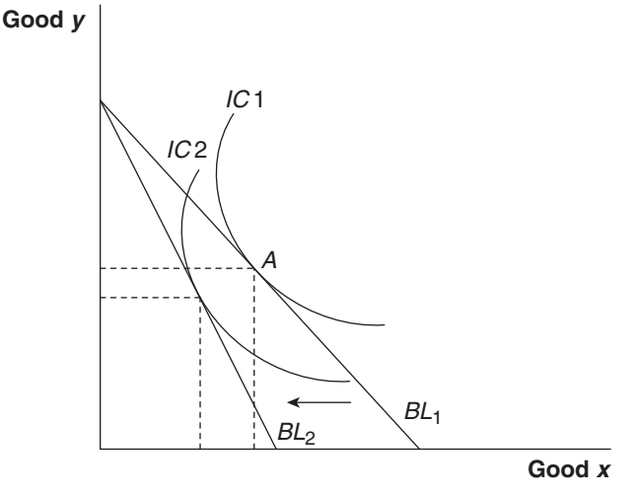


**Figure 49.2** Consumer Choice Problem

If, however, we broaden our notion of utility, the purchase of fair trade is rational. First, consider decision utility as the satisfaction derived from the outcome itself. Given a choice between A and B, if A is chosen, then A has been revealed as preferred over B. On the other hand, diagnostic utility refers to the utility that an individual gains from his or her actions. So for example, the choice to donate to a worthy cause may increase self-esteem and provide additional utility. The total utility received is the sum of the decision utility and the diagnostic utility. Losses in decision utility due to overpaying for a good or experience can be offset by the increase in diagnostic utility by participating in what is viewed as a worthy cause. Consumers of fair trade coffee receive utility from the consumption of the coffee itself and the satisfaction of participating in a worthy cause.

Now consider the coffee supply decision. Coffee producers must decide on whether to participate in fair trade coffee production. As discussed above, fair trade standards set a minimum price for coffee. This creates a price floor ($P_{ft}$) and excess supply. As shown in Figure 49.3, the fair trade price is above the equilibrium price ($P_E$), and supply ($Q_2$) exceeds demand ($Q_1$).

The higher price guaranteed by fair trade creates an incentive to increase production beyond the market clearing quantity. Because fair trade is a relatively small share of the total coffee market, many coffee farmers can sell only a portion of their output as fair trade, with the remainder spilling into the conventional market and sold at prevailing prices.

Fair trade, however, raises production costs in complying with certification standards. Coffee producers therefore must make a decision on the expected profits from fair trade production based on their costs and expected demand. As long the total premium earned for the volume of coffee sold in the fair trade market exceeds the extra costs of production, there is an incentive to produce fair

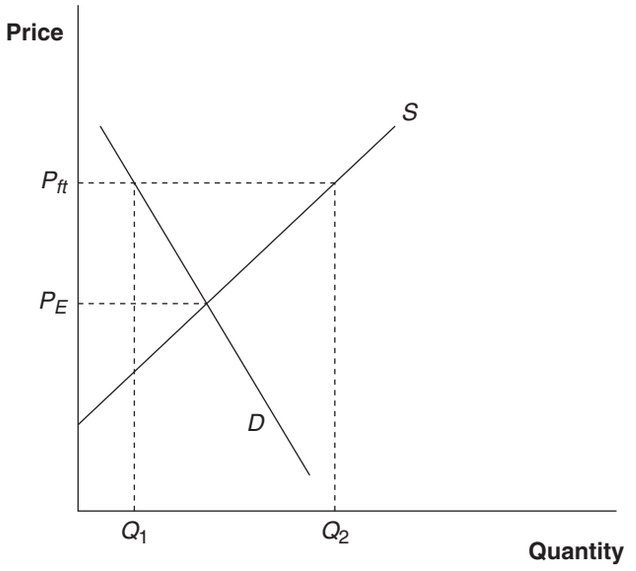


**Figure 49.3** Fair Trade as a Price Floor

trade. With weak demand for fair trade, excess production is sold in the conventional market, often at prices below cost, leading to losses for farmers. Therefore, fair trade producers still face income uncertainty.

An alternate way to think about fair trade is in relation to market failures, specifically uncompetitive markets. Value chain analysis looks at how profits are distributed along the way from ripe fruit to coffee in the cup. As coffee moves through the distribution chain from grower to consumer, it passes through several hands and is transformed in some manner at each stage. If markets are competitive, each stage of production will receive profits equal to the value added at their respective stage. However, when markets are not competitive, individuals or firms can use their market power to extract surplus value at the expense of someone else along the value chain.

Talbot (1997) provides an instructive analysis of the coffee production chain. He divides the process into three main components: growing/initial processing on the farm, processing up to the green coffee stage and exporting, and importing green coffee/final processing and sale of roasted or instant coffee to consumers.

Talbot (1997) calculates the share of the total sale price of a pound of ground roasted coffee that accrues along the various stages of the value chain from 1971 to 1995. This includes several periods of booms and busts in the coffee market, most notably the Brazil frost of 1975 that devastated the 1976 coffee crop. His results indicate that the share of the per pound price of ground roasted coffee going to the grower has been falling from an average of 21.4% in the 1970s to 19.5% in the 1980s and 13.3% in the 1990s. Moreover, over the time frame Talbot analyzed, the retail price of coffee rose in 13 years, but the price paid to the growers in dollar terms fell in 5 of those 13 years. A driving force in the decline of income paid to coffee growers from the mid-1980s was the unraveling of the International Coffee Agreement in 1986.

The breakdown of the ICA allowed a handful of multinational corporations to effectively hold down the price of green coffee beans and raise the price of processed coffee in consuming countries. By controlling green coffee exports, companies such as Kraft, Nestlé, Procter & Gamble, Sara Lee, and Tchibo can use their market power to keep prices low. Historically, coffee producers have had little access to markets and pricing information and no distribution channel, leaving them at the mercy of large multinational vendors.

Fair trade attempts to address this market failure. While fair trade mandates a minimum price for growers, it also creates more direct distribution between consumers and producers. The elimination of several layers in the distribution chain raises grower earnings. Furthermore, the social premium allows communities to invest in infrastructure, reducing production costs and strengthening direct market contact. Pooling coffee growers into cooperatives creates economies of scale and increases bargaining power.

Value chain analysis convinced many that markets are biased against the world's poor. Structural adjustment policies and later actions by the World Trade Organization created a market backlash and brought attention to alternate trade arrangements. The arrival of fair trade became an important market-based pro-poor initiative. The popular mantra of "trade, not aid" signifies that fair trade is not a handout but a trade alternative that promotes respect and dignity as it works to alleviate poverty.

## Evidence

While fair trade advocacy groups provide anecdotal evidence highlighting the success of fair trade, there is limited research into the relative effectiveness of fair trade in achieving its goals. In a seminal paper, Raynolds, Murray, and Taylor (2004) consider the gains of fair trade coffee to producers, cooperatives, and communities. They analyzed seven fair trade coffee cooperatives: five in Mexico, one in Guatemala, and one in El Salvador. They report that households in the cooperatives have higher and more stable income. Fair trade participation generated two to three times the earnings of conventional sales for households. The authors suggest that while many conventional producers were forced to migrate to cities when coffee prices fell below production costs, fair trade members were able to stay on their land. Furthermore, they were able to use the premium from fair trade to invest in safe water and education. The premium also allowed diversification of farm activities, enhancing income and security.

Cooperatives also provide benefits to their members. Part of the fair trade premium was used by cooperatives to build infrastructure, increase storage capacity, and invest in processing plants. This increases local processing, lowers

transportation costs, and enhances profits for members. Cooperatives have also used a portion of the fair trade premium to provide medical services to the community and frequently provide additional social benefits such as education, housing, and even stores selling low-priced necessities.

Bacon, Ernesto Méndez, Eugenia Flores, and Brown (2008) conducted a household-level survey on 177 coffee-producing households in northern Nicaragua. The sample contained 101 households that are part of a fair trade cooperative, 61 that sell in conventional markets, and 15 that sell certified organic coffee. Their results are framed within the United Nations Millennium Development Goals, of which only two will be discussed.

They estimate that the average household profits from the production of coffee were about $971 per year. Based on average household size, this corresponds to about $0.38 per person per day, well below the United Nation threshold for extreme poverty of $1. Furthermore, there was no significant earnings difference between fair trade, conventional, or organic producers. Income from coffee sales, whether fair trade or conventional, was insufficient to eradicate extreme poverty in this sample, with 69% of the households reporting they were sometimes unable to meet basic nutritional needs.

There were, however, significant educational differences between fair trade and conventional households. Fair trade households sent 97% of 7- to 12-year-olds to primary school, above the rate for conventional coffee households (74%) and the national average (88%). The difference continued into secondary school, where 84% of 13- to 17-year-olds in fair trade households were attending compared to 53% from conventional households.

Overall, the results from Bacon et al.'s (2008) research are mixed. While it appears that participation in fair trade led to a greater commitment to education, there were no significant differences in earnings associated with fair trade production. Further income generated from any type of coffee production was insufficient to lift households out of extreme poverty and resulted in food insecurity.

In a noncoffee study, Becchetti and Costantino (2008) consider the impact of fair trade on herb growers in Kenya. They focus on Meru Herbs, a commercial organization created in 1991 by an association of local farmers as a means to fund canals on the Kitheno River. Roughly 97% of Meru Herbs sales are from fair trade exports of both organic and conventional products. Meru Herbs signs contracts with farmers who either have organic certification or are in the process of obtaining it. The farmers agree to sell some of their produce to Meru Herbs and in exchange receive complimentary seeds, organic fertilizer, low-cost fruit trees, training courses, and technical assistance. Meru Herbs also buys fruit from conventional farmers with no additional benefits provided.

The authors conducted a survey and divided the respondents into four distinct groups: organic producers, conversion producers, fruit-only producers, and a control group. The organic producers are certified and have a signed contract with Meru Herbs. Conversion producers have signed a contract but have not completed the conversion to organic production. Fruit-only producers do not have a contract but sell fruit to Meru. The control group consists of farmers from the same region who have access to the canal project but no connection to fair trade outlets. The benefits of fair trade are measured using income, health, child labor, and education.

The groups affiliated with fair trade all reported higher income than the control group. The premium for fair trade ranged from 65% for conversion producers to 38% for fruit-only producers. It is interesting that the conversion producers have higher monthly earnings than established organic producers. One explanation may be that a premium is paid to lure the producers to convert to organic. It may also be that only the best nonorganic producers are given contracts, creating potential selection biases.

The results from the health measures are less clear. Groups affiliated with fair trade reported lower child mortality rates than the control group, with child mortality rates ranging 42% to 75% lower. There was virtually no difference in child vaccination rates between fair trade and the control group. Furthermore, while organic and conversion farmers were more likely to have their last child born in a hospital, the fruit-only and control groups were equally likely at 60%.

Moving to child labor and education, we again see mixed results. The authors measure child labor based on the percentage of children attending school. The control group reported that on average, 23% of household children between the ages of 6 and 15 were attending school, while 19% of teens between 15 and 18 were attending school. This is better than organic producers (13% and 9%, respectively) and fruit-only producers (8% and 4%, respectively). Only conversion producers reported higher rates than the control group (45% and 25%). This may provide further support for selection bias among those who convert from traditional to organic production.

These results support the notion that fair trade raises income, with fair trade–affiliated producers receiving on average 53% more in monthly earnings than the control group. The impact on health, child labor, and education is less clear. The control group reports higher child mortality rates than fair trade producers but lower levels of child labor and greater commitment to education than two of the fair trade groups.

## Future Directions

It is clear that fair trade has stirred great interest and passion among citizens of developed countries as a method to improve well-being among impoverished producers. The question remains as to whether fair trade can achieve this goal. Empirical results to date have been mixed, and more research needs to be done to understand the impact of fair trade on producers. The focus needs to expand beyond income and look at broader issues such as vulnerability and security to see if fair trade has the capacity to improve well-being.

A promising avenue for fair trade to explore is boutique coffee. Increasing quality standards may lead to appreciable taste differences between fair trade and conventional coffee, increasing demand. There is also some evidence that producers gain from direct market access, but more needs to be done in this area as well. Marketing and distributing coffee requires expertise and generates additional costs. Distributors must have sufficient resources to finance coffee transactions. If indeed direct market access promotes farmers' earnings, then investment in infrastructure and outreach will expand the connection between producers and consumers.

## Conclusion

Fair trade is a market-based pro-poor trade alternative that offers trade, not aid. It arose from the backdrop of structural adjustment eliminating many social programs for the world's poor and the collapse of the International Coffee Agreement. The outcome of several social justice campaigns, fair trade promises higher and more stable income, sustainable development, and respect and dignity for working poor people around the world. Consumers participate in fair trade to express their preferences for social justice and solidarity for producers. Coffee growers incur higher costs to produce in accordance with fair trade standards and do so to generate higher earnings. With historical market failures in coffee production due to the market power of multinational coffee vendors, fair trade is a way of breaking down barriers between producers and consumers, allowing direct distribution of coffee. The impact of fair trade on well-being is mixed. Some results suggest that fair trade leads to higher earnings, while others indicate no significant difference in earnings for fair trade and conventional producers.

## References and Further Readings


Bacon, C. (2005). Confronting the coffee crisis: Can fair trade, organic, and specialty coffees reduce small-scale farmer vulnerability in northern Nicaragua? *World Development, 33,* 497–511.

Bacon, C., Ernesto Méndez, V., Eugenia Flores, M., & Brown, M. (2008). *Will "we" achieve the millennium development goals with small-scale coffee growers and their cooperatives? A case study evaluating fair trade and organic coffee networks in northern Nicaragua.* Santa Cruz: University of California, Santa Cruz, Center for Agroecology & Sustainable Food Systems Research Briefs.

Barratt Brown, M. (1993). *Fair trade: Reform and realities in the international trading system.* London: Zed Books.

Becchetti, L., & Adriani F. (2002). *Fair trade: A "third generation welfare" mechanism to make globalisation sustainable* (Working Paper 170). Rome: University of Rome Tor Vergata, Centre for International Studies on Economic Growth.

Becchetti, L., & Costantino, M. (2008). The effects of fair trade on affiliated producers: An impact analysis on Kenyan farmers. *World Development, 36,* 823–842.

Bodner, R., & Prelec, D. (1997). *The diagnostic value of actions in a self-signaling model.* Retrieved from http://www.cepr.org/meets/wkcn/3/3503/Papers/Prelec.pdf

Clement, S., & Defranceschi, P. (2006). *BUY FAIR: A guide to the public purchasing of fair trade products.* Toronto, Ontario, Canada: ICLEI—Local Governments for Sustainability.

Fridell, G. (2004). The fair trade network in historical perspective. *Canadian Journal of Development Studies, 25,* 411–428.

Jaffee, D. (2007). *Brewing justice: Fair trade coffee, sustainability and survival.* Berkeley: University of California Press.

LeClair, M. (2002). Fighting the tide: Alternative trade organisations in the era of global free trade. *World Development, 30,* 949–958.

Maseland, R., & de Vaal, A. (2002). How fair is fair trade? *De Economist, 150,* 251–272.

Murray, D., Raynolds, L. T., & Taylor, P. L. (2003). *One cup at a time: Fair trade and poverty alleviation in Latin America.* Retrieved August 8, 2008, from http://www.colostate.edu/Depts/Sociology/FairTradeResearchGroup

Murray, D., Raynolds, L. T., & Taylor, P. L. (2004). Building producer capacity via global networks. *Journal of International Development, 16,* 1109–1121.

Ponte, S. (2002). The "Latte Revolution"? Regulation, markets and consumption in the global coffee chain. *World Development, 30,* 1099–1122.

Prebisch, R. (1971). *Change and development: Latin America's great task; report submitted to the Inter-American Development Bank.* New York: Praeger.

Raynolds, L. (2002). *Poverty alleviation through participation in fair trade coffee networks: Existing research and critical issues.* Background paper prepared for project funded by the Community and Resource Development Program, the Ford Foundation, New York. Retrieved August 8, 2008, from http://www.colostate.edu/Depts/Sociology/FairTradeResearchGroup

Raynolds, L., Murray, D., & Taylor, P. (2004). Fair trade coffee: Building producer capacity via global networks. *Journal of International Development, 16,* 1109–1121.

Renard, M. (1999). The interstices of globalisation: The example of fair coffee. *Sociologia Ruralis, 39,* 484–500.

Simpson, C., & Rapone, A. (2000). Community development from the ground up: Social-justice coffee. *Human Ecology Review, 7,* 46–57.

Steinrucken, T., & Jaenichen, S. (2006). Does the fair trade concept work? An economic analysis of social labels. *Aussenwirtshaft, 61,* 189–209.

Talbot, J. (1997). Where does your coffee dollar go? The division of income and surplus along the coffee commodity chain. *Studies in Comparative International Development, 32,* 56–91.

U.S. Department of Agriculture (USDA). (1999). *Agricultural outlook.* Available at http://www.ers.usda.gov/publications

Waridel, L. (2002). *Coffee with pleasure: Just java and world trade.* Montreal, Quebec, Canada: Black Rose Books.

Weber, J. (2007). Fair trade coffee enthusiasts should confront reality. *Cato Journal, 27,* 109–117.

Wrigley, G. (1988). *Coffee.* London: Longman.

Yanchus, D., & de Vanssay, X. (2003). The myth of fair prices: A graphical analysis. *Journal of Economic Education, 34,* 235–240.

Zadek, S., & Tiffen, P. (1996). Fair trade: Business or campaign? *Development, 3,* 48–53.

# PART VI

## Economic Analyses of Issues and Markets

# 50

# Economics of Education

Ping Ching Winnie Chan
*Statistics Canada*

Economics is a study of making choices to allocate scarce resources—what to produce, how to produce, and for whom to produce. With education being one of the top social priorities to compete for scarce resources, economic analysis on returns of education, the optimal allocation of education resources, and evaluation of education policies has become increasingly important. In addition to the theoretical development and empirical applications of human capital theory, different areas of economics have also applied to education issues, such as the use of the economics of production to study the relation between education inputs and outputs.

Since the release of the 1983 National Commission on Excellence in Education report *A Nation at Risk*, the U.S. government has initiated different education reforms to improve failing student performance. In recent years, rich data sets (often longitudinal) have become available from different education policy experiments—for example, voucher programs in Milwaukee and accountability policies in different states. Research that takes advantage of these policy experiments has shed some light in the current education debate by providing empirical evidence of the impact of different policies and suggestions for policy design.

In this chapter, the human capital theory and an alternative screening hypothesis are first introduced to provide some theoretical background in the area. Next, the impact of various education inputs on student performance is discussed to apply the production theory on education and to highlight some of the empirical challenges in evaluating the impact of key factors in education. Then, theoretical and empirical evidence on school choice and accountability policies in the elementary and secondary school markets is described to highlight the policy relevance of the field.

## Theory

### Human Capital Theory

Education is often regarded as one of the main determinants in one's labor market success. The primary model used to measure the returns to education in the field is the human capital theory. Drawing an analogy to the investment theory of firms, each worker is paid up to his or her marginal productivity—the increment of output to a firm when an additional unit of labor input is employed. Human capital theory postulates that education is made as human capital investment to improve one's productivity and earnings.

The essence underlying human capital theory can be traced back to Adam Smith's classic theory of equalizing differentials, in which the wages paid to workers should compensate for various job characteristics. The modern human capital theory, building on the seminal work of Theodore Schultz (1963), Gary Becker (1964), and Jacob Mincer (1974), has drawn attention to conceptualizing the benefits of education and to modeling education as a form of investment in human capital.

As a first step to calculating the rate of returns to education, Schultz (1963) categorized a list of education benefits, including the private benefits in terms of wages and the social benefits that are accrued to the economy as a whole (e.g., research and development to sustain economic growth and better citizenship). In calculating the costs of education, in addition to the direct costs of tuition, transportation, books, and supplies, the indirect costs of forgone earnings—the opportunity costs incurred while acquiring the human capital during schooling—should also be included.

Introduced by Becker (1964), one way to determine the optimal level of education is to compare the internal rate of

return ($i$) earned by acquiring that level of education and the market rate of interest ($r$) in financing to acquire that level of education. The internal rate of returns to education corresponds to the implicit rate of return earned by an individual acquiring that amount of education, and the market rate of interest represents the opportunity cost of financing the investment required to obtain the amount of education. Because of diminishing returns, the present value of marginal benefits of education generally declines with years of education. The present value of marginal costs (direct cost and the opportunity cost of forgone earnings), on the other hand, generally increases with years of education. The combined actions of marginal benefits and marginal costs of education result in a falling internal rate of return when education attainment rises. As a simple illustration, assuming a constant cost of borrowing in the capital market, Figure 50.1 shows that one will continue to invest in education as long as the internal rate of return ($i$) exceeds the market rate of interest ($r$) and that the optimal equilibrium education (E*) level has been attained when $i = r$.

## Estimating the Rate of Returns to Education

To empirically estimate the rate of returns to education, cross-sectional data on earnings of different individuals at a point in time can be compared, and the differences in earnings can be attributed to differences in the levels of education after accounting for other observable differences. Mincer (1974) developed a framework to estimate the rate of returns to education in the form of the human capital earnings function as follows:

$$\log y_i = a + rS + bX + cX^2 + e, \quad (1)$$

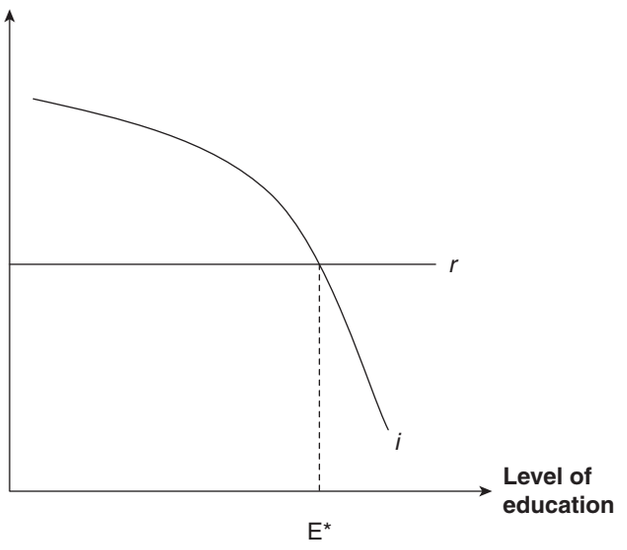


**Figure 50.1** Optimal Education Level Determination

NOTE: $i$ is the rate of return; $r$ is the market interest rate; E* is the optimal level of education.

where $y$ are the earnings of individual $i$. $S$ represents the years of education acquired, $X$ represents individual's potential experience that is usually approximated by $X = \text{age} - S - 6$, and $e$ is the stochastic term.

The Mincer (1974) model is widely used to estimate returns of schooling, and the estimate $r$ can be interpreted as the rate of returns to education when assuming that each additional year of education has the same effect on earnings and the years of schooling are measured accurately. Using the U.S. Census data, James Heckman, Lance Lochner, and Petra Todd (2006) showed that the data from recent U.S. labor markets do not support the assumptions required for interpreting the coefficient on schooling as the rate of returns to education; the review calls into question using the Mincer (1974) equation to estimate rates of returns to education. To better fit the data in modern labor markets, many studies have extensions and modifications of the original Mincer equation—for example, allowing wage premiums when one completes elementary school in Grade 8, high school in Grade 12, and college after 16 years of completed education (the sheepskin effect), as well as incorporating other interaction terms or nonlinearities into the earnings equation.

A growing literature has also attempted to measure the causal effect of schooling by addressing the concern of ability bias in standard earnings equations. By assuming that, because of their same genetic makeup, identical twins' innate ability would be the same, Orley Ashenfelter and Cecilia Rouse (1998) estimated the rate of returns to education on a large sample of identical twins from the Princeton Twins Survey. Their results confirm that there is ability bias in the standard estimation, and once the innate ability is controlled for, the estimate of the returns to education is lowered.

## Screening

The human capital theory emphasizes the role of education in enhancing one's productivity. An alternative model based on the work of Kenneth Arrow (1973) and Michael Spence (1973) argued that education plays a role as a signal of workers' productivity when productivity is unknown before hiring because of imperfect information. The assumption is that if the individuals with higher ability have lower costs of acquiring education, they are more likely to acquire a higher level of schooling. In such a case, the level of education one possesses can act as a signal to employers of one's productivity, and individuals will choose an education level to signal their productivity to potential employers.

It is empirically difficult to distinguish between the human capital theory and the screening hypothesis because high-ability individuals will choose to acquire more education under either theory. In an attempt to contrast the two models, Kevin Lang and David Kropp (1986) examined U.S. enrollment data and compulsory attendance laws from

1910 to 1970. Under the human capital theory, the effect of an increase in minimum school-leaving age should be apparent only among individuals that have optimal years of education lower than the minimum school-leaving age. Under the signaling model, the increase in minimum school-leaving age will also affect individuals that are not directly affected under the policy in the human capital theory. The reason is that a rise of the minimum school-leaving age from $t$ to $t + 1$ would decrease the average ability level of individuals with $t + 1$ years of education. To signal higher ability, the high-ability individuals will choose to stay in school for $t + 2$ years. Therefore, the human capital theory and the signaling theory would yield different implications for the changes in compulsory attendance laws. Lang and Kropp's analysis showed that enrollment rates did increase beyond those directly affected by the rise in minimum school-leaving age

Thus far, empirical evidence in the area is still limited, and the results have not been conclusive in determining the role of the two theories in education investment decisions. On one hand, the pure signaling model cannot explain the investment decision in professional education programs such as medical science or law because it is only through the professional programs that one can accumulate the knowledge and skills necessary for those occupations; on the other hand, it is impossible to discount the value of signaling completely from the empirical data. Given the important role of education in our society, both theories contribute in explaining the motivation of education investment decisions, and the current literature suggests that the importance of each theory might differ by the level and program in which an individual enrolls.

## Education Production Function

In the production theory, a production function typically indicates the numerical function by which levels of inputs are translated into a level of output. Similarly, an education production function is usually a function mapping quantities of measured inputs in schooling (e.g., school resources, student ability, and family characteristics) to some measure of school output, like students' test scores or graduation rates. A general education production function can be written as follows:

$$T_{it} = f(S_{it}, A_i, F_{it}, P_{it}), \tag{2}$$

where $T_{it}$ is the measured school output, such as test scores of student $i$ at time $t$. For the inputs in the production function, $S$ measures resources received by students at schools that are affected by school expenditure, teacher quality, class size, as well as curriculum planning and other school policies. $A$ denotes the innate ability of a student, $F$ measures the family background characteristics that are related to student education attainment, and $P$ measures peer effects from classmates and close peers.

Many empirical studies have attempted to measure the impact of school resources on performance because school inputs can be more directly affected by policies targeted to improving student performance. One of the earliest studies in the area is the landmark 1966 *Equality of Educational Opportunity* report led by James S. Coleman. Aiming to document the different quality of education received by different populations in the United States in response to the 1964 Civil Rights Act, the author collected detailed information on schools, teachers, and students. One of the most controversial conclusions from the report is that measurable characteristics of schools and teachers played a smaller role in explaining the variation of student performance compared to family characteristics. Though the Coleman report's interpretation has been criticized from the regression framework (e.g., in Eric Hanushek and John Kain, 1972), the report remains a benchmark study in assessing education outcomes.

In a widely-cited paper, Hanushek (1986) surveyed the literature on the impact of school inputs and concludes that there is no evidence that real classroom resources (including teacher–pupil ratio, teacher education, and teacher experience) and financial aggregates (including teacher salary and expenditure per pupil) has a systematic impact in raising student performance. Reviewing research with new data and new methodologies in a follow-up, Hanushek (2006) maintained that the conclusion about the general inefficiency of resource usage in public schools is still warranted; further, he suggested that more research is needed in the area to understand when and where resources are most productively used.

### Class Size

A popular policy proposal is to reduce class size to improve student performance. With smaller class sizes, students can have more personal interactions with teachers during class time. However, as pointed out by Caroline Hoxby (2000), measuring the impact of class size has been difficult because the greater part of class size variation documented in administrative data "is the result of choice made by parents, schooling providers, or courts and legislatures" (p. 1240). The observed variations in class sizes, therefore, are correlated with some unobservable factors that might be correlated with student performance.

To ascertain the effect of class size reduction, the Tennessee Department of Education conducted a 4-year longitudinal class size study, the Student Teacher Achievement Ratio (STAR). The cohort of students who entered kindergarten in the 1985 to 1986 school year participated in the experiment through third grade. A total of 11,600 children were involved in the experiment over all 4 years. The students were randomly assigned into one of three interventions: small class (13 to 17 students per teacher),

regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students per teacher with a full-time teacher's aide). Classroom teachers were also randomly assigned to the classes they would teach. The class size cutoffs were designed following the study conducted by Gene Glass and Mary Smith (1979). They summarized the findings on the performance impact of class size from 77 studies comparing student performance in different class sizes and concluded that the greatest gains in achievement occurred among students who were taught in classes of 15 students or fewer.

The STAR project provides an opportunity to study the causal effect of class size given that the students are assigned to different class size randomly to avoid the endogeneity problem of the class size variable. Alan Krueger (1999) controlled for observable student characteristics and family background and found that relative to the standard deviation of the average percentile score, students in smaller classes performed better than those in the regular classes (both with and without an extra teaching aide). The effect sizes are 0.20 in kindergarten, 0.28 in first grade, 0.22 in second grade, and 0.19 in third grade, which suggest that the main benefit of small class size tended to concentrate by the end of the first year. In light of this finding, Krueger suggested that small class size "may confer a one-time, 'school socialization effect' which permanently raises the level of student achievement without greatly affecting the trajectory" (p. 529).

Using an alternative measurement approach, Hoxby (2000) exploited the natural variation in population size from a long panel of enrollment data in Connecticut school districts. Similar to Project STAR, the class size variation that resulted from cohort size variation is exogenous in the estimation because class sizes are not driven by choices of parents or the school authority. However, using this natural population variation in class size, the analysis found no significant impact of class size on student performance.

## Family Factors

In addition to school resources, innate ability is an important contributor in student performance that is often unobservable in data analysis. To further complicate the analysis, family background, especially parental education, is often correlated with unobserved student ability. While reviewing the literature on school resources and student outcomes, David Card and Krueger (1996) pointed out that the presence of omitted variables might bias the relation between school resources and student outcomes because children from better family backgrounds are likely to attend schools with more resources.

Having some family controls in the empirical analysis is important to control for the influence family factors might have on student outcomes. Nonetheless, there has been limited evidence identifying the impact of family factors on student performance separately. Take family income as an example: Most studies can show only a suggestive correlation between student outcomes and family income but not of a causal relationship between family income and student outcome (Mayer, 1997). Family income is likely to suffer from the problem of endogeneity under such a reduced-form regression model because children growing up in poor families are likely to face other adverse challenges that would continue to affect their development even if family income were to increase. To address the issue of endogeneity, different papers have attempted to isolate the exogenous variation of income using different research designs. Gordon Dahl and Lance Lochner (2008) exploited the income shocks generated by the large, nonlinear changes in the Earned Income Tax Credit (EITC) in the past two decades. Their analysis found a positive impact of income on student mathematics and reading scores, and the effects were stronger for children of younger ages. Exploiting the policy variation in child benefit income among Canadian provinces, Kevin Milligan and Mark Stabile (2009) also showed that child benefit programs in Canada had significant positive effects on different measures of child outcomes. More research in the area is still needed to understand the impact of family factors and to prescribe relevant policy recommendations.

## Peer Effects

A distinct feature of the education production function is that there is a public good component in student performance. Classroom teaching simultaneously provides benefits to more than one student at the same time, and the learning experience of one student is being affected by the average quality of his or her classmates (peer effects). Edward Lazear (2001) explicitly modeled the negative externalities created from the disruptive behavior of one student on all other classmates and provided the theoretical intuition to reconcile the mixed results documented in the class size reduction analysis. In particular, the model implied that class size effects are more pronounced in smaller classes for students with special needs or from disadvantaged populations.

Empirical evidence in identifying peer effects, however, has been rather limited. Charles Manski (1993) pointed out that analyses from standard approaches that regress students' outcomes on peer outcomes cannot properly isolate peer effects from the selection effect, in which it is difficult to separate a group's influence on an individual's outcome from the individual's influence on the group—the reflection problem. To put it simply, the reflection problem occurs when the performance of Student A is influenced by the presence of Student B, and vice versa. Another challenge in identifying peer effects empirically is that peers are rarely assigned randomly. Parents are likely to select neighborhoods through residential location to associate with good peers in good neighborhoods with good schools. A positive correlation between peer and individual might

reflect this unobserved self-selection instead of any direct impact of peers on student performance.

The effect of increased sorting from increased school choice initiatives is often discussed in the choice policy debate. When more choice is given to parents, students remaining in the public school will suffer if the better students choose to leave the public school and switch to the alternative and thereby lower the peer quality remaining in the public school. The impact of sorting based on outcomes is discussed in the school choice section.

## Education Policy

To improve the efficiency and effectiveness of public schools, a variety of reform strategies have been proposed on the policy front. These include restructuring the school organization and decision-making process, improving accountability, and providing more choice among schools as reflected in policies such as open enrollment programs that increase choice in the public school system or private school vouchers and tuition tax credits between public and private schools. In this section, the policy implications and empirical evidence on school choice and accountability measures are discussed to highlight the policy relevance of economic analysis in education.

### School Choice

The promotion of choice among schools through the use of financing mechanisms such as vouchers and tuition tax credits has a long history. Milton Friedman's original essay on the government's role in education, published in *Capitalism and Freedom* (1962, chap. 6), is often regarded as the source that reinvigorated the modern voucher movement. In the essay, Friedman advocated the use of government support for parents who chose to send their children to private schools by paying them the equivalent of the estimated cost of a public school education. One of the key advantages of such an arrangement, he argued, would be to permit competition to develop and stimulate improvement of all schools.

The concepts of choice and competition are closely related and tend to be used interchangeably in this area of research. Yet, as defined in Patrick Bayer and Robert McMillan (2005), there is a clear distinction between the two, with choice relating to the availability of schooling alternatives for households and competition measuring the degree of market power enjoyed by a school. Typically, one would expect a positive correlation between the two, with an increase in choice being associated with an increase in competition. According to the standard argument, increased competition will force public schools to improve efficiency in order to retain enrollment. Yet in principle, increased choice need not always improve school performance, as illustrated in a model developed by McMillan (2005). His model showed that rent-seeking public schools could offset the losses from reduced enrollment by cutting costly effort.

Using public money in the form of vouchers or tax credits to support private schools is perhaps one of the most contentious debates in education policy today. School choice advocates claim that increasing school choice could introduce more competition to the public system to improve overall school quality and increase the access to alternative schools for low-income students, minority students, or both, who would otherwise be constrained to stay in the public system. Opponents argue that it is inappropriate to divert funds from the already cash-strapped public system because it would have a detrimental impact on school quality. According to the opponents' view, voucher or tax credit policy is likely to benefit only affluent families who could afford the additional cost of attending private schools given that the vouchers and tax credits are unlikely to cover the total cost of attending private schools.

Despite the controversy of school choice debate, there are now a number of U.S. states implementing various forms of school choice programs: Arizona, Florida, Illinois, Iowa, Minnesota, Pennsylvania, and Puerto Rico have adopted some forms of tax credit, deductions, or both; Colorado, the District of Columbia, Florida, Ohio, and Wisconsin have publicly funded school voucher programs in place; and Maine, Vermont, and New York have voucher programs funded by private organizations.

Empirical studies that examine the effect of private schools as well as different choice programs are discussed to highlight the important role of economic analysis in evaluating policy programs and the challenge in identifying the school choice impact.

#### *The Effects of Private Schools*

One strand of the choice literature has focused on public versus private provision, to see whether student outcomes differ across sectors, controlling for student type. In practice, students self-select into public or private schools, and this nonrandom selection might bias the performance effect in a comparison of outcomes between public and private schools since private school students may be unobservably better.

In an early study, Coleman, Thomas Hoffer, and Sally Kilgore (1982) used data from the High School and Beyond Survey of 1980 and showed that Catholic schools were more effective than public schools. Jeffrey Grogger and Derek Neal (2000) used the National Educational Longitudinal Survey of 1988, which allowed analysis of secondary school outcomes conditional on detailed measures of student characteristics and achievement at the end of elementary school. They found achievement gains in terms of high school graduation rates and college attendance from Catholic schooling, and the effects were much larger among urban minorities. Given the difficulty in controlling for selection bias in the measurement approach,

only suggestive evidence on positive private school effects can be drawn from these studies.

*Small-Scale Randomized Voucher: The Milwaukee Experiment*

Wisconsin was the first U.S. state to implement a school voucher program, the Milwaukee Parental Choice Program, to low-income students to attend nonreligious private schools in 1990. Only students whose family income was at or below 1.75 times the national poverty line were eligible to apply for the vouchers. The number of successful applicants was restricted to 1% of the Milwaukee public school enrollment in the beginning of the program and was later raised to 1.5% in 1994 and 15% in 1997. One thing worth noting in the Milwaukee program is that only nonreligious private schools are allowed to participate in the program.

Using the normal curve equivalent reading and math scores from the Iowa Test of Basic Skills (ITBS), which was administered to the Choice students every year and in Grades 2, 5, and 7 for students in the Milwaukee public schools, Rouse (1998) compared the test scores of Choice students with those of unsuccessful applicants and other public school students. The sample consisted of African American and Hispanic students who first applied in the years 1990 to 1993 for the program, and a matching cohort was drawn from the public student data with valid test scores for comparison. Rouse's results showed that the voucher program generated gains in math scores, and students selected for the voucher program scored approximately two extra percentage points per year in math when compared with unsuccessful applicants and the sample of other public school students. In contrast, there was no significant difference in reading scores.

*Tuition Tax Credit: Ontario, Canada*

In 2001, the provincial government of Ontario, the most populous province in Canada, passed the Equity in Education Tax Credit Act (Bill 45), which led to an exogenous increase in choice between public and private schools. The tax credit applied to the first $7,000 of eligible annual net private school tuition fees paid per student. Although 10% of eligible tuition fees could be claimed for the tax year 2002, a 50% credit rate was scheduled to be phased in over a 5-year period, resulting in a maximum tax credit of $3,500 when the program was fully implemented in 2007, which was a comparable amount to that offered in the Milwaukee experiment (about $3,200 in 1994 to 1995). The tax credit was large in scope because about 80% of private schools participated in the program, irrespective of their religious orientation. In 2003, the credit was switched off unexpectedly because of a change in the provincial political party in power.

Taking advantage of this unique opportunity to study a large-scale increased choice experiment in North America, Ping Ching Winnie Chan (2009) examined the Grade 3 standardized test (an annual assessment administered by the Education Quality and Accountability Office in Ontario) of Ontario public schools before and after the introduction of the tax credit policy. The empirical evidence indicates that the impact of competition was positive and significant in 2002 to 2003 (the year the tax credit was in full effect). The average proportion of students attaining the provincial standard was found to be about two percentage points higher in districts with greater private school presence—districts where competition from private schools would be higher because of the tax credit. The Ontario experience provides evidence that increasing private school choice could help public schools to improve their performance.

*Choice Within the Public System*

In addition to an increase in school choice from private schools, there are other forms of choice within the public system—for example, open enrollment policies, magnet schools, and charter schools. Students in North America typically attend a neighborhood public school assigned within the school attendance boundary, and the application to out-of-boundary schools varies from jurisdiction to jurisdiction. Open enrollment policies allow a student to transfer to another school either within or outside his or her school attendance area. This form of choice is indeed the most prominent form of choice available to students in the public system. Magnet schools are schools operated under the public system, but they exist outside of the zoned boundaries and offer academic or other programs that are different from the regular public schools to attract enrollment. Charter schools, on the other hand, are not subjected to the administrative policies within the public system even though they are also publicly funded. The charter documents the school's special purpose and its rules of operation and will be evaluated by the declared objectives in the charter. In the United States, there are about 200 charter schools, mostly situated in Minnesota and California, and in Canada, Alberta is the only province with charter schools. The magnet schools and the charter schools aim to provide more diversity within the public system to meet the needs of students.

*Open Enrollment Policy: Chicago Public Schools*

In Chicago, students have more flexibility in their high school selection and may apply to schools other than their neighborhood schools when space is available through the Chicago Public Schools (CPS) open enrollment policy. The take-up rate of the policy among Chicago students is quite high, and more than half of the students in CPS opt out of their neighborhood schools.

Using detailed student-level panel data from CPS, Julie Cullen, Brian Jacob, and Steven Levitt (2005) compared the outcomes of those who did versus those who did not

exercise choice under the open enrollment program. Students opting out of their assigned high schools (i.e., those exercising choice) were found to be more likely to graduate compared to those who remained in their assigned schools. An important insight in their study, though, is the presence of potential selection bias in the sample of students who take advantage of the choice program. If the more motivated students are more likely to opt out of local schools and to fare better academically than students with lower motivation, then opting out may be correlated with higher completion rate. This calls into question whether to interpret the higher education outcomes as an effect from the school choice program itself or from the selection bias in motivation or other unobservable characteristics among the students who opted out.

As mentioned, one of the important issues in the school choice literature relates to the performance impact from increased student sorting. Although competition varies, the mix of students between public schools and alternative schools would also vary. Thus, in a more competitive environment, the average ability of students in public schools may be systematically different from students in public schools operating in a less competitive environment. When school choice increases, productivity responses from schools and sorting effects due to changes in the mix of students are likely to operate together. It is an empirical challenge to separate the effects of sorting from productivity responses, and more research in the area is still required in this direction.

## Accountability

Much of the earlier debate in U.S. educational policy for raising performance concentrates on providing more resources to the system, such as increasing educational expenditures, reducing pupil–teacher ratios, or supporting special programs to target students with different needs. As performance responses to these policy initiatives have shown to be less than encouraging (see Hanushek, 1986), federal policy since the 1990s has moved to emphasize performance standards and assessments. Since the passage of the No Child Left Behind Act (NCLB) in 2001, all states are required to follow standards and requirements laid down by the Department of Education to assess the annual performance of their schools. The policy has been controversial, and debates have focused on its effectiveness in improving student performance as well as other unintended outcomes—for example, grade inflation, teaching-to-the-test, or gaming the system, which might result when the system responds to specific objective measures. Working with the CPS, Jacob and Levitt (2004) presented an investigation of the prevalence of cheating by school personnel in the annual ITBS. The sample included all students in Grades 3 to 7 for the years from 1993 to 2000. By identifying "unusually large increases followed by small gains or even declines in test scores the next year and unexpected patterns in students' answers" (p. 72) as indicators of cheating and a retesting experiment in 2002, the authors concluded that cheating on standardized tests occurs in about 4% to 5% of elementary school classrooms. The frequency of cheating was also found to increase following the introduction of high-stakes testing, especially in the low-performing classrooms. The investigation showed that cheating is not likely to be a serious problem in high-stakes testing but confirmed that unintended behavioral responses would be induced under a different incentive structure. This creates a challenge for policy makers to design incentive structures that could minimize the disruptive behavior. In this section, an overview of the NCLB legislation and empirical evidence on student performance impact from stronger accountability measures in Florida and Chicago are discussed.

### *The No Child Left Behind Legislation*

Immediately after President George W. Bush took office in 2001, he proposed the legislation of NCLB, aiming to improve the performance of U.S. elementary and secondary schools. The act requires all states to test public school students in Grades 3 through 8 annually and in Grades 10 to 12 at least once, subject to parameters set by the U.S. Department of Education. The states set their own proficiency standards, known as *adequate yearly progress* (AYP), as well as a schedule of target levels for the percentage of proficient students at the school level. Such performance-based assessment reinforces the accountability movements that were adopted in many states prior to NCLB (by 2000, 39 states had some sort of accountability system at the school level). If a school persistently fails to meet performance expectations, it will face increasing sanctions that include providing eligible students with the option of moving to a better public school or the opportunity to receive free tutoring, which the act refers to as *supplemental educational services.* States and school districts also have more flexibility in the use of federal education funds to allocate resources for their particular needs, such as hiring new teachers or improving teacher training.

### *Florida's A+ Plan for Education*

Florida's A+ Plan for Education, initiated in 1999, implemented annual curriculum-based testing of all students in Grades 3 through 10. The Stanford-9 Achievement Test (SAT-9) was instituted as the Florida Comprehensive Assessment Test (FCAT) in 2000. All public and charter schools receive annual grades from A to F based on aggregate test performance, and rewards are given to high-performing schools, while sanctions as well as additional assistance are given to low-performing schools. The most controversial component of the plan is the opportunity scholarships, which are offered to students attending schools that received an F grade for 2 years during a 4-year period. In 2006, the Florida Supreme Court issued a ruling declaring the private school option of the Opportunity Scholarship Program unconstitutional, and students are no

longer offered the opportunity to transfer to a private school; however, the option to attend a higher-performing public school remains in effect.

Prior to the A+ Plan, in 1996 Florida had begun rating schools based on their aggregate test performance (based on a nationally norm-referenced test such as the ITBS or SAT-8) on students in Grades 3 through 10 in reading and mathematics. Based on student performance, schools were stratified into four categories on the Critically Low Performing Schools List, and schools receiving the lowest classification were likely to receive a stigma effect because of the grading.

David Figlio and Rouse (2006) examined the Florida student-level data from a subset of school districts from 1995 to 2000 to assess the effects of both the voucher threat and the grade stigma on public school performance. The analysis showed positive effects on student outcomes in mathematics in low-performing schools, but the primary improvements were concentrated in the high-stakes grades (prior to 2001 to 2002, students in Grades 4, 8, and 10 were tested in reading and writing, and students in Grades 5, 8, and 10 were tested in math). Their results also found that schools concentrated more effort on low-performing students in the sample.

### *Chicago Public Schools*

In 1996, Chicago Public Schools (CPS) introduced a school accountability system in which elementary schools were put on probation if the proficiency counts (i.e., the number of students who achieved a given level of proficiency) in reading on the ITBS were lower than the national norm. Schools on probation were required to implement improvement plans and would face sanctions and public reports if their students' scores did not improve.

After the introduction of NCLB in 2002, CPS designated the use of proficiency counts based on the Illinois Standards Achievement Test (ISAT), which had been introduced by the Illinois State Board of Education in 1998 to 1999 as the statewide assessment, as the key measure of school performance. Subject to the requirement of the act, the ISAT test in 2002 changed from a relatively low-stakes state assessment to a high-stakes exam under the national assessment.

Derek Neal and Diane Schanzenbach (in press) examined data from CPS to assess the impact of the introduction of these two separate accountability systems. The analysis compared students who took a specific high-stake exam under a new accountability system with students who took the same exam under low stakes in the year before the accountability system was implemented. The results show that students in the middle of the distribution scored significantly higher than expected. In contrast, students in the bottom of the distribution did not score higher following the introduction of accountability, and only mixed results could be found among the most advantaged students.

Both the Florida and the Chicago results show performance effects on public school performance after the introduction of stronger accountability measures. However, given that the effects are not uniformly distributed across students with different abilities and are often concentrated in students of high-stakes grades, both studies raise concerns about resource allocation when responding to such accountability measures and call for more research on policy design that can minimize disrupted incentives.

## Conclusion and Future Directions

The economics of education is a growing and exciting field in economic research. As reviewed in this chapter, the empirical analysis in the field has already provided many insightful ideas for current education issues that are relevant to public policy. The new research in the area has started to adopt a more structural approach to understand the impact of education reform policies that have often induced responses from different players—for example, using general equilibrium analysis to model the linkage between housing decisions and school consumption to measure the potential benefits on school quality with school choice policy (see Bayer, Fernando, & McMillan, 2007; Nechyba, 2000). These new papers use innovative measurement design and more detailed microlevel data and can complement the findings from reduced-form analysis in the literature as well as identify the channels through which the policy effect is being brought about.

## References and Further Readings

Arrow, K. (1973). Higher education as a filter. *Journal of Public Economics, 2*(3), 193–216.

Ashenfelter, O., & Rouse, C. (1998). Income, schooling, and ability: Evidence from a new sample of identical twins. *Quarterly Journal of Economics, 113*(1), 253–284.

Bayer, P., & Fernando, F., & McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy, 115,* 588–638.

Bayer, P., & McMillan, R. (2005). *Choice and competition in local education markets* (Working Paper No. 11802). Cambridge, MA: National Bureau of Economic Research.

Becker, G. (1964). *Human capital.* New York: Columbia University Press.

Belfield, C. (2000). *Economic principles for education.* Cheltenham, UK: Edward Elgar.

Benjamin, D., Gunderson, M., Limieux, T., & Riddell, C. (2007). *Labour market economics: Theory, evidence, and policy in Canada* (6th ed.). Toronto, Ontario, Canada: McGraw-Hill Ryerson.

Card, D., & Krueger, A. (1996). School resources and student outcomes: An overview of the literature and new evidence from North and South Carolina. *Journal of Economic Perspective, 10*(4), 31–50.

Chan, P. (2009). *School choice, competition, and public school performance.* Unpublished doctoral dissertation, University of Toronto, Ontario, Canada.

Checchi, D. (2006). *The economics of education.* Cambridge, MA: Cambridge University Press.

Cohn, E., & Geske, G. (1990). *The economics of education* (3rd ed.). Oxford, UK: Pergamon.

Coleman, J. (1966). *Equality of educational opportunity.* Washington, DC: Government Printing Office.

Coleman, J., Hoffer, T., & Kilgore, S. (1982). *High school achievement: Public, Catholic and private schools compared.* New York: Basic Books.

Cullen, J., Jacob, B., & Levitt, S. (2005). The impact of school choice on student outcomes: An analysis of the Chicago Public Schools. *Journal of Public Economics, 89,* 729–760.

Dahl, G., & Lochner, L. (2008). *The impact of family income on child achievement: Evidence from the earned income tax* (Working Paper No. 14599). Cambridge, MA: National Bureau of Economic Research.

Figlio, D., & Rouse, C. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics, 90,* 239–255.

Friedman, M. (1962). *Capitalism and freedom.* Chicago: University of Chicago Press.

Glass, G., & Smith, M. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis, 1*(1), 2–16.

Grogger, J., & Neal, D. (2000). Further evidence on the effects of Catholic secondary schooling. *Brookings-Wharton Papers on Urban Affairs, 1,* 151–193.

Hanushek, E. (1986). The economics of schooling. *Journal of Economic Literature, 24,* 1141–1177.

Hanushek, E. (2006). School resources. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 2, pp. 865–908). New York: Elsevier North-Holland.

Hanushek, E., & Kain, J. (1972). On the value of "equality of educational opportunity" as a guide to public policy. In F. Mosteller & D. P. Moynihan (Eds.), *On equality of educational opportunity* (pp. 116–145). New York: Random House.

Hanushek, E., & Welch, F. (Eds.). (2006). *Handbook of the economics of education* (Vols. 1 & 2). New York: Elsevier North-Holland.

Heckman,J., Lochner, L., & Todd, P. (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 1, pp. 307–458). New York: Elsevier North-Holland.

Hoxby, C. (2000). The effect of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics, 115*(4), 1239–1285.

Krueger, A. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics, 114*(2), 497–532.

Jacob, B., & Levitt, S. (2004, winter). To catch a cheat. *Education Next, 4*(1), 69–75.

Lang, K., & Kropp, D. (1986). Human capital versus sorting: The effects of compulsory attendance laws. *Quarterly Journal of Economics, 101*(2), 609–624.

Lazear, E. (2001). Educational production. *Quarterly Journal of Economics, 116*(3), 777–803.

Levin, H. (1989). Mapping the economics of education: An introductory essay. *Educational Researcher, 18*(4), 13–16, 73.

Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies, 60,* 531–542.

Mayer, S. (1997). *What money can't buy: Family income and children's life chances.* Cambridge, MA: Harvard University Press.

McMillan, R. (2005). Competition, incentives, and public school production. *Journal of Public Economics, 89,* 1131–1154.

Milligan, K., & Stabile, M. (2009). *Do child benefits affect the wellbeing of children? Evidence from Canadian child benefit expansions* (Working Paper No. 12). Vancouver, British Columbia, Canada: Canadian Labour Market and Skills Researcher Network.

Mincer, J. (1974). *Schooling, experience, and earnings.* New York: Columbia University Press.

Neal, D., & Schanzenbach, D. (in press). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics.*

Nechyba, T. (2000). Mobility, targeting, and private school vouchers. *American Economic Review, 90*(1), 130–146.

Rouse, C. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics, 113*(2), 553–602.

Schultz, T. (1963). *The economic value of education.* New York: Columbia University Press.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics, 87*(3), 355–374.

# 51

# Economics and Justice

George F. DeMartino
*University of Denver*

Exploration of the connections between economics and justice is complicated by the fact that both economics and justice are variously defined and deeply contested. Economics can be thought of as the study of the determinants of wealth creation, which was the view taken by many classical economists; the study of choice and allocation under conditions of scarcity, which is the view of contemporary neoclassical theory; the study of the production, appropriation, and distribution of the social surplus, which informs Marxian theory; the study of the manner in which communities provision for themselves, which is the view of many institutionalists; and in other ways. Justice is likewise a controversial concept. Here, too, one finds that distinct approaches yield diverse accounts of what is just and unjust. Justice can be defined in terms of fairness, respect for inviolable rights, or equality. Making matters more interesting, each of these definitions can be (and is) also theorized in diverse ways. For instance, egalitarians (those who define justice in terms of equal distribution) often disagree among themselves about whether a just distribution is one that assures equality of income, wealth, opportunities, rights, or something else entirely. Finally, many economists tend to avoid discussion of ethical matters, including justice, which they define as lying outside their field of expertise. As a consequence, notions of justice are implicit rather than explicit in much of economics.

This diversity implies that it is a mistake to think that there is any one right way to conceptualize economics, justice, or the connections between them. It is both intellectually honest and far more rewarding to embrace the conceptual diversity that one encounters in this context and then to explore the ways in which distinct approaches to economics engage distinct notions of justice. This allows for an informed investigation of conceptions of economics and justice. One stands to learn much about the controversies among economists over policy and the assessment of economic outcomes by attending to the justice conceptions that inform their judgments. In a reciprocal manner, one can better evaluate contending justice claims by attending to their economic implications. After all, an approach to justice that seems entirely plausible in the abstract may generate concern if one finds that its economic policy implications are damaging or even repugnant in fundamental ways.

This chapter proceeds in just this way. It begins with and gives most attention to neoclassical economic theory—the orthodox approach to economics that has predominated in the profession for many decades. One encounters a deep ambivalence among neoclassical economists about the place of moral judgments in economics—including but not limited to judgments concerning justice. This suggests that there is no one neoclassical conception of justice. This chapter instead investigates the theoretical strategies that have emerged within the tradition to manage this ambivalence. Despite the reluctance of neoclassical economists to engage matters of morality, a conception of justice has emerged within important strands of neoclassical thought. It is associated with what are called *negative* liberties, rights, and freedoms. This is an approach that focuses on the individual's freedom from illegitimate constraints on his or her decision making. This approach informs the work of Nobel Laureate Milton Friedman and other economists. A particularly clear statement of this approach appears in the work of libertarian philosopher Robert Nozick (1974), and the chapter examines his central arguments and their implications for economic justice.

Then the chapter turns to one heterodox approach to political economy, Marxian theory, and explores alternative notions of justice that inform this approach. Marxian theory is exemplary of many other heterodox approaches that embrace positive rights and freedoms. These reach far beyond the requirements associated with negative rights. This chapter explores two relevant aspects of Marxian theory: (1) its notion of exploitation and (2) its conception of a just distribution of burdens and rewards. These are derived from its normative grounding in an egalitarian notion of justice.

It bears emphasis that the controversy in economics over justice is not just of academic interest. The debate is deeply consequential. To demonstrate this, this chapter investigates what each approach implies about the legitimacy of the market as an institution for distributing income, wealth, and opportunity. This is one of the most important of all policy debates in economics and politics. One sees that what a school of economic thought has to say about this question depends directly on the conception of justice to which it is committed.

## Neoclassical Economics

There is a profound difficulty within neoclassical theory in speaking about justice. This difficulty stems from the conception of economics as value-free science that informs neoclassical thought. Advocates of neoclassical theory tend to view economics as an objective science, the goal of which is to theorize the functioning of economic processes and outcomes as they are, undistorted by normative judgments. Just as physics seeks to describe the physical world as it is, not as the physicist would like it to be, so must the economic scientist describe the operation of the social world independent of normative biases. Neoclassical economics holds to a rigid distinction between positive economics—the objective explanation of economic phenomena—and normative economics—the evaluation of economic outcomes and the formulation of policy prescription. In this dichotomy, positive economics is by far preeminent. Considerations of justice reside in the domain of normative economics and are taken to be far less scientific than the considerations that inform positive economics. As a consequence, most neoclassical literature is altogether silent on justice concerns.

Neoclassical theory incorporates the effort to remain value free in the specification of the assumptions with which it begins its work. Human actors are presumed to be rational, where rationality is defined in particular ways. For neoclassical theory, rationality implies self-interested, egoistic behavior. Rational actors seek their own welfare. Moreover, they always desire more of the things they like. Rationality also implies that they know best what is good for them. This requires an assumption that each agent has a preference ordering that maps his or her full set of likes, dislikes, values, commitments, tolerances, intolerances, fears, and passions. When confronted with a choice between two commodities or courses of action, rational agents simply consult their preference ordering and make those choices that benefit them most, all things considered. By *benefits most* is meant maximizing their personal happiness or utility, or more prosaically, simply choosing those options that best satisfy their preferences.

The assumption of rationality yields an important implication. In this framework, there is no basis for the economist or the economic theory to judge the choices that economic actors make. Because they are rational, they know best, full stop. It is of critical importance that this theoretical strategy spares the economist from having to make the value judgments that assessing actors' preferences would require. Hence, the assumption of rationality allows economics to undertake its investigation of economic matters in an apparently objective way.

Rational agents have virtually limitless desires, but they populate a world of finitude. This is because all output requires inputs from nature, but nature's bounty is finite. At any one moment, only so much can be produced. Hence, economic actors confront a world of scarcity. This leads to the central economic problem in neoclassical thought: choice and allocation under conditions of scarcity. How can the scarce resources that society has available to it be best allocated across the diverse purposes to which they can be put? Should more energy go into auto production or into mass transportation? Which choice will leave society best off, where best off is theorized in terms of maximizing people's satisfaction of desires, according to their own respective preference orderings?

It would seem that the neoclassical emphasis on scarcity and allocation of resources would immediately call forth questions of economics of justice, because in a world of scarcity there is not enough to go around. There would seem to be a need here for normative criteria to assess the justness of any particular distribution of total social output. For reasons already discussed, however, this is not the case. Neoclassical theory's commitment to objective science requires that it attempt to answer questions surrounding allocation of scarce resources and distribution of final output while minimizing value judgments. Shunning value judgments has led the profession to emphasize efficiency as its chief evaluative criterion rather than alternative, value-laden criteria like justice. In this approach, an economic outcome is taken to be inefficient if at least one person could be made better off given existing resources and technologies without making at least one other person worse off. For example, if one person prefers apples to bananas but has in his or her possession only bananas, while another person has the opposite preference but possesses only apples, the situation is inefficient because both could be made better off simply by exchanging with each other. In contrast, an efficient (or Pareto optimal) outcome is one in which no one can be made better off without making at

least one other person worse off. If one assumes that these two people do trade bananas and apples, the situation that results after the trade is made is efficient.

Notice that the assessment of efficiency does not seem to require of the economist any value judgments. The economist does not ask or care why one person prefers apples and another bananas. The economist does not ask how many apples or bananas each possesses prior to or after the exchange takes place. That is taken to be none of the economist's business. And so it appears that judgments pertaining to efficiency are largely value free.

## Neoclassical Economics, Distribution, and Justice

What does this approach suggest about distribution of income and wealth? Efficiency considerations imply that income should be distributed in whatever way best ensures the enhancement of social welfare. That distribution is best that permits the maximum satisfaction of preferences, given existing resources and technologies. This distribution can be summed up as follows: In the first instance, to each according to his her contribution.

Most neoclassical economists emphasize the efficiency benefits of rewarding each agent according to his or her contribution (rather than the intrinsic rightness of such arrangements). The efficiency argument is simple and is understood by its advocates to entail little in the way of value judgments. The argument is this: An economic system that rewards agents for their contributions will give them an incentive to contribute more—perhaps by investing in training (and so enhancing their human capital) or by acquiring other productive assets like land or capital. The assumption of rationality implies that individuals do this for themselves, in pursuit of increasing income, but the unintended consequence of their doing so is increasing total social output owing to their increasing productivity. Hence, the argument for rewarding agents according to their contributions is, for most economists, a simple matter of wise economic management. Economies that reward contribution will grow prosperous, and that will allow for greater satisfaction of people's desires.

But what about distributive justice as opposed to efficiency? A central theme in neoclassical thought is that this is largely a noneconomic question, because it entails value judgments of the sort that neoclassical theory does not permit. For example, imagine that a professor distributes $100 among all his or her students, equally. Is this arrangement efficient? The answer is yes, because once the distribution is made, no one can be made better off without making another worse off (the only way for one student to get more is for another to get less). But what if, instead, a professor gives all $100 to just one student? Is this distribution efficient? The answer is again yes because the only way to make any one of the unfortunate students who received nothing better off is to take some amount of money away from the lucky student who possesses the $100. Both situations are equal in an efficiency sense. This implies that efficiency has nothing to do with equity, fairness, or justice. And because neoclassical thought emphasizes efficiency over other evaluative criteria, one is led to the conclusion that there is little more to be said about them.

### *Free Choice, Negative Freedom, and Justice*

There is more to the story, however. Neoclassical theory's attachment to rationality entails a strong commitment to individual free choice, which in turn implies a certain sense of justice that some (though by no means all) leading neoclassical economists embrace. Because the agents know best what is in their best interest, that economic system is best that allows them the freedom to choose from among the opportunities they face. There are constraints in this choosing, of course. As discussed, choices are constrained by scarcity. In a world of scarcity, each and every action taken entails an opportunity cost in terms of what must be forgone to pursue it. The material that goes into a bicycle tire is unavailable for producing an automobile tire. In a market economy, scarcity is imposed on each agent via his or her budget constraint. Individuals can purchase whatever they choose, provided they do not exceed the resources they have available to them (through income or borrowing).

This sense of economic freedom is often described as *negative freedom* and is associated with negative rights or liberties (Berlin, 1958). Negative freedom entails the right to choose from among the opportunities one confronts, given one's circumstances (such as one's budget), without coercion by others (and especially by government). In this account, a poor person and a rich person may be equally free, provided that each can choose without interference from among the opportunities each confronts. Naturally, the rich person will have a greater set of opportunities than the poor person. But if both can choose freely from among the options available to them, each is equally free in the sense of the enjoyment of negative rights.

This manner of thinking then leads to a particular conception of justice. A distribution is taken to be just, provided that it arises from the exercise of free choice defined as negative freedom. Agents are to be free to make the best deals available to them within the constraints set by their budgets. Whatever aggregate distribution of social wealth arises from each agent acting in this manner is on this account just, because it arises from the equal enjoyment of negative freedom of all agents, no matter how rich or poor they might be. In this account, an unjust distribution would be one that arose from the violation of some people's (negative) rights. For instance, a distribution that arises from theft, extortion, or physical coercion (such as through the threat of force) would be unjust because it violates persons' negative rights.

This conception of justice emerges in the work of some important neoclassical economists who engage matters of

justice. Friedman (1962; Friedman & Friedman, 1990) is particularly notable in this regard. In work such as *Capitalism and Freedom* and *Free to Choose,* he advocates forcefully for negative rights in the form of freedom of choice as restricted only by one's budget constraint. For Friedman, the enjoyment of such rights is vital to economic prosperity, for the reasons already discussed—and even to justice. He argues strenuously against physical coercion in economic affairs, especially by the state, as inherently unjust. He equates freedom with the absence of such coercion.

The equation of justice with the outcome of free choice derives from the work of political theorists like John Locke (1690) and is developed more fully in contemporary libertarian work. Locke argues that people have an inherent right to own that which they create—both directly through their own labors, and indirectly through the capital they employ to hire the labor of others. Locke reaches this important normative conclusion on the basis of a simple intuitive argument. First, Locke contends that "every Man has a *Property* in his own *Person.* This no Body has any Right to but himself" (MacPherson, 1962, p. 200). This is a natural right that cannot be legitimately abridged by society or government. Second, people have a right to attempt to survive, and this can be accomplished only by their appropriating for themselves elements of nature. One achieves this by mixing labor with nature, thereby transforming it into necessary goods. Hence, though nature is provided to humankind in common, the act of individual labor provides the normative foundation for individual appropriation. Each person acquires the exclusive right to possess and consume that which his or her own labor has created. In Locke's words, "The *Labour* of his Body, and the *Work* of his Hands, we may say, are properly his" (MacPherson, p. 200).

Locke extends this right of property to include that output produced by the labor of others whom one has hired. He reaches this conclusion by emphasizing that the right of property in one's own person and labor must entail the right to alienate it to others through voluntary exchange. In this respect, one's labor is no different from other property.

The modern libertarian view of distribution, such as that advocated by prominent libertarian theorist Nozick, is consistent with these Lockean propositions. In this view, any distribution of a society's output is just, provided that it arises only from legitimate processes. In Nozick's (1974) words, "The complete principle of distributive justice would say simply that a distribution is just if it arises from another (just) distribution by legitimate means" (p. 151). Presuming a just prior distribution—that is, one that did not arise via the infringement of people's rights—the outcome of a series of voluntary exchanges between free individuals must also be deemed just, regardless of the patterns of distribution that arise as a consequence. If those who make the greatest contribution are able to secure through voluntary exchange a greater reward, then the consequent inequality is entirely just. Indeed, from this perspective, government initiatives to redistribute income in pursuit of greater equality are unjust. In Nozick's view, because such initiatives essentially force some to work, uncompensated, for others, redistributive measures are even tantamount to slavery.

Just one matter remains. Tying reward to contribution (or voluntary exchange) speaks to distribution in the first instance. But what of those who do not contribute or who have nothing to exchange? Many in society produce nothing at all, either during portions of their lives (when they are infants, elderly, or infirm), or during their entire lives (if they are in some way incapacitated in ways that prevent so-called productive work). Few would argue for their complete exclusion from a share of total output: Beneficence toward the deserving poor is widely accepted among advocates of most theoretical perspectives.

This begs the question: What degree of provisioning should society make for the unproductive? Here, neoclassical theory provides little guidance because this question is seen to lie squarely in the domain of value judgments. Instrumental (efficiency) arguments still apply, of course: In cases where the unproductive can be made productive, the argument can be made that a sufficient distribution should be made to induce this rehabilitation. In this case, value neutrality might be (apparently) preserved via the recommendation of the use of cost-benefit analysis as the appropriate scientific means for making this judgment. But this then would be seen as the nonnormative matter of determining just what distribution to the unproductive will generate an efficient outcome. Beyond this, neoclassical theory has little to say.

## Egalitarianism and Heterodox Economics

Many alternative approaches to economics and political economy differ from neoclassical theory in their initial assumptions, substantive propositions, analytical methods, and especially in their normative judgments. Not least, approaches as diverse as Marxian, institutionalist, socioeconomic, and feminist theory tend toward the embrace of egalitarian normative commitments, as do many contributions to political theory (see Anderson, 1990; Lutz, 1999; Tool, 1979; Walzer, 1973, 1983). These are commitments that emphasize equality among society's members. As Amartya Sen (1992) has argued at length, however, most normative frameworks emphasize the equality of something that is taken to be fundamental. Distinct frameworks differ primarily in what it is that each seeks to equalize across society's members. And so it should not be surprising that distinct heterodox approaches tend to define what it is that is to be equalized differently. Indeed, even within each of these traditions, we find normative controversy including, but not limited to, what makes for a just outcome.

The Marxian approach is representative of the many diverse schools of thought that embrace some form of egalitarianism, and so this chapter investigates it in some detail here. Like other heterodox approaches, the Marxian tradition reaches beyond negative freedom (associated with negative rights) to embrace positive freedom (tied as it is to positive rights). Positive freedom (sometimes called *substantive freedom*) concerns not just the absence of constraints on a person's actions. Instead, according to Sen (1992) it speaks to the full range of beings and doings that a person can actually achieve or enjoy, given his or her income, wealth, race, gender, level of schooling, and all other factors that bear on what a person can be or do. This account recognizes that two individuals with equal negative freedom (in the sense of freedom from coercion) may enjoy very different levels of positive freedom. The relatively poor person who is entirely free to choose may face a very bleak opportunity set as compared with that of a rich person. Egalitarian frameworks that privilege positive as opposed to negative freedom are inclined to find ethically deficient a social arrangement that fails to address this inequality. They seek reform that expands the opportunity set of the disadvantaged in order to equalize the life chances (and not just the negative rights) of society's members.

John Rawls (1971) has been particularly influential in shifting the attention of political theorists and others to positive conceptions of freedom. In *A Theory of Justice,* Rawls advances an approach to justice as fairness. He investigates what kind of social arrangements concerning distribution (and other things) would arise voluntarily from a process of rational deliberation among society's members. He asks us to join him in a thought experiment in which we suppose a committee of deliberators who will decide on the best institutional arrangements for the society that they will inhabit. Critically, he requires that one envisions this committee as doing its work behind a veil of ignorance—that is, they must design the rules under which they and all other members of society will live, without knowing as they deliberate into which group in society they themselves will be placed. This ensures that they will consider the fairness of the arrangements they propose from the perspective of all the groups that make up society. Rawls argues that the outcome of this thought experiment would be just, because it would potentially be deemed fair by all of society's members.

What principles would such a committee of deliberators agree on to govern distribution in their society? Rawls (1971) argues that the committee would settle on two fundamental principles. The first requires the equal distribution of primary goods to all of society's members. Primary goods are the "basic rights, liberties and opportunities, and the . . . all-purpose means such as income and wealth," but also the "bases of self-respect." These goods, Rawls continues, "are things citizens need as free and equal persons" (pp. 180–181). Justice as fairness requires that these goods be equally provided to all of society's members so that each has equal substantive ability (positive freedom) to pursue his or her life plans.

Rawls's (1971) second principle, the difference principle, modifies the first. It allows for the case in which the equal distribution of primary goods may harm all of society's members. Justice as fairness permits inequality in the distribution of primary goods, provided that the worst off benefit most thereby. This test for inequality is quite demanding: It requires evidence that unequal distributions help most those who will receive least. Absent such evidence, Rawls argues that inequality in primary goods is illegitimate.

Rawls's work has been deeply influential. In economics, Sen has taken up and extended Rawls's work in ways that relate to economic conceptions of justice (see also Nussbaum, 1992). Sen argues that while Rawls is right to emphasize equality of positive freedom, he errs by focusing on the means to achieve freedom rather than on the actual freedom that people enjoy. This is because individuals differ in their abilities to convert primary goods into achievements. This implies that two people with equal bundles of primary goods may nevertheless face distinct levels of substantive freedom. For instance, a disabled person may need greater income and support than others to achieve the same level of beings and doings (such as mobility or occupational success). Were all individuals identical, this problem would not arise. But because interpersonal differences are so dramatic, the goal of promoting equality in substantive freedom requires that we focus on substantive equality directly, rather than on the means to achieve it. And this implies that we may need to distribute primary goods and other means to achieve unequally if we are to achieve the most important kind of equality: equality in the positive freedoms that people enjoy.

## Positive Freedom and Marxian Economics

Look at how this emphasis on positive freedom bears on normative judgments within the Marxian framework. This framework begins with a simple, intuitively plausible assumption: To exist over time, all societies must produce a surplus. That is, those who perform the labor necessary for provisioning (producing food, clothing, health care, shelter, etc.) must produce not just enough to sustain themselves, but to sustain others. This is because at any particular moment, some in society will be unable to produce for themselves. This is true of infants, young children, the elderly, the infirm, and the otherwise disabled. If the productive workers produced only enough to sustain themselves, society would be unable to reproduce itself over time.

The Marxian framework concerns itself principally with the diverse ways that societies organize the production, appropriation, and distribution of the surplus (Resnick & Wolff, 1987). Who is assigned to produce the

social surplus, and what means are used to ensure that they are induced to perform this social necessity? Who is entitled to appropriate the surplus so produced? That is, who receives it, and by what juridical, political, cultural, or other means are they ensured that the surplus flows to them? Finally, what norms, rules, laws, or conventions dictate the distribution of the social surplus across society's members, including those who participate in and those who are excluded from the practices of producing and appropriating the social surplus?

Unlike neoclassical theory, many heterodox approaches explicitly engage normative questions and even base their theoretical frameworks on normative judgments (DeMartino, 2000). This is certainly true of the Marxian approach. Vital questions arise immediately in this approach about the rightness of the arrangements that societies adopt to manage the production, appropriation, and distribution of the social surplus. First, one finds in the Marxian tradition a normative indictment of what is called *exploitation.* This term is used to describe any social arrangement in which those who produce the surplus are excluded from its appropriation. This happens whenever others in society enjoy the right of appropriation at the expense of the producers. For instance, in slave societies, the surplus is produced by slaves but appropriated by the slave owners. This, for Marxists, represents a particularly clear example of exploitation. On this ground (among others), slavery is deemed unjust. And so would be any other social arrangement that shared the feature of exploitation.

The Marxian tradition has much to say not just about who should appropriate the surplus, but also about how that surplus should ultimately be distributed (DeMartino, 2003; Geras, 1985, 1992; Lukes, 1987). In this approach, a distribution of the social surplus is just when it is based on need. This conception is consistent with Sen's framework, as already discussed. Those who require more, perhaps because of physical disabilities (permanent or temporary), age, geography, or other challenges, are entitled to greater shares of the social surplus. In contrast, distributive justice requires that those who require the least because of good fortune or other factors are to receive less (Marx, 1938).

Distribution according to need relates directly to the positive freedom that underlies the Marxian approach. Allocating more to those with the greatest need ensures that they enjoy increasing substantive freedom to live valued lives. It expands their opportunity sets by increasing the beings and doings that are available to them. Distributing more to those who are most impoverished in terms of their freedoms and who face the greatest challenges thereby contributes toward the equalization of positive freedom across society's members. It generates what its advocates see as genuine equality. This is equality in positive freedom—in what people can actually achieve—rather than in what this approach's proponents view as the hollow freedom associated with neoclassical theory's emphasis on negative freedom.

The equality that underpins Marxian and other heterodox approaches is much more demanding than is the equality to which neoclassical theory is committed. The latter requires only the removal of certain kinds of constraints—especially those imposed by the state, given its monopoly over the legitimate use of violence to enforce its dictates.

For those who value positive freedom, the removal of these constraints may not be at all sufficient to the achievement of genuine equality. Indeed, the value placed on negative rights might often interfere with the realization of equality of positive freedom. This would be the case, for instance, when the exercise of rights by those who are best off interferes with and reduces the opportunity sets of those who are less advantaged. Hence, we find a tension between these two kinds of freedoms and especially between demands for equality of the one as opposed to the equality of the other.

## Applications and Policy Implications

Normative judgments matter deeply—not only at the level of abstract debate, but also in the design and evaluation of institutions, policies, and outcomes. Indeed, the positions economists take on the most important public policy questions are tied directly to the normative frameworks (including their judgments about justice) that underlie their approaches to economics. To see this, consider the question of whether the economic outcomes associated with the market economy are just. Here one finds a sharp controversy between those who embrace negative and those who embrace positive accounts of freedom.

In Friedman's (1962) account, distribution of total social wealth under the free market results from free exchange in the marketplace. Free exchange here entails only that there is no coercion. In the absence of such coercion, agents are taken to be entirely free to pursue their own interests as they see fit. This ensures that they will enter into only those agreements (or undertake those market exchanges) that improve their personal welfare. It is not for the economist to evaluate a person's judgments in this regard. When someone consummates an exchange, economists must presume that that person has made the best bargain available to him or her.

This conception of market interactions and outcomes implies that agents will receive rewards in the marketplace that are commensurate with their contributions to social welfare. Those who have produced the goods that society deems most useful or desirable will secure a higher price when they sell these goods than will other agents who are selling less desirable goods. Moreover, those with the greatest skills and with savings that they are willing to invest will make the greatest contributions to total output, which in turn serves as the means to enhance social welfare. In a free market, these agents will be able to bargain for rewards that reflect these greater contributions. In contrast, those with

few skills or other endowments will receive low rewards that are commensurate with their lower contributions to total output. This is as it should be: Justice prevails when agents secure the share of total income they are due owing to what they contribute and the choices they make, unconstrained by illegitimate infringements on their rights or the rights of others.

From this perspective, the free market is viewed as the optimal form of economic arrangement. Not only is it apt to promote economic efficiency owing to the incentives it provides each actor to make greater contributions, but it also is just in the sense that it ensures equal negative rights to all actors. Unlike other economic arrangements that are based on dictates from the state that necessarily compromise negative freedom, the marketplace operates on a logic of freedom that allows each agent to pursue his or her self-interest unimpeded by illegitimate interference. So it is that Friedman can equate capitalism with freedom.

Those who embrace positive, as opposed to negative, freedom reach a radically different conclusion about free market processes and outcomes. The Marxian tradition argues that the negative rights associated with the marketplace both obscure and deepen substantive inequality. In a capitalist economy, those who produce the social surplus, the wage laborers, are exploited by the firms that hire them because the legal right to claim the surplus produced by laborers is monopolized by the firm. The surplus is extracted from workers unfairly, despite the illusion of fairness given by the fact that under capitalism workers are free to work or not as they see fit (Bowles & Gintis, 1990; Resnick & Wolff, 1987; Roemer, 1988). In this sense, the apparently free workers under capitalism are no different from slaves because both are deprived of the right to appropriate their own surplus. Hence, Marx can call employment under capitalism wage slavery.

From the Marxian perspective, there are other reasons to deem the free market unjust. The market economy does not ensure that those with the greatest need will secure the greatest allocations of the social surplus. Instead, those who are worst off in terms of their needs will often be least able to bargain for a fair price for what they have to sell. This is particularly the case for workers. Because under capitalism the means of production are monopolized by capitalists, workers cannot sustain themselves independently in the ways that they could were they to have in their possession the means of production. They are therefore compelled to sell their labor power in a market that is stacked in favor of the capitalists. Hence, free exchange leads to systematic unfairness owing to the asymmetry in bargaining power between capital and labor.

This concern about the inherent injustice of the labor market is shared by other heterodox traditions, such as radical institutionalism (see Dugger, 1989; Dugger & Sherman, 1994). For instance, institutionalist economist John Commons (1924) argues that workers must secure a wage regularly in order to survive: Because what they have to sell is perishable, they are forced to take what they can get (Ramstad, 1987). In contrast, the firms to which they must sell their labor power often can wait. This asymmetry in the ability to wait leads to an asymmetry in bargaining power, which in turn ensures that wages will be depressed below the fair level that would exist were workers and firms to confront each other as equals. For Commons, the outcome of such asymmetric bargaining is on its face unjust. In his view, reasonable value in exchange is realized only when both parties enjoy equal ability to wait.

## Future Directions

Much works remains to be done on the connections between economics and justice. At present, neoclassical theory is undergoing substantial change owing to new avenues of research. For instance, developments in behavioral and experimental economics are demonstrating the severe limitations of the rationality assumption. Research indicates that individuals are often driven in their decision making by notions of fairness, justice, the welfare of others (not just family members but even strangers), and other ethical concerns. This implies new avenues for research on several fronts, as Sen (1987) has argued. What conceptions of justice do individuals hold when they make decisions? How do these conceptions affect their behavior? How are these conceptions (and behavior) affected by the social milieus in which individuals act? And what kinds of institutional arrangements might be desirable in promoting the conceptions of justice that individuals value? Although the answers to these questions will require substantial research, it is increasingly apparent that the simplistic account of rationality that has informed neoclassical economics for many decades is unlikely to survive much longer in economics. This may encourage neoclassical economists to revisit their historical antipathy toward normative judgments in their work.

Heterodox economic research can and likely will be extended to encompass a greater focus on normative matters and the kinds of institutions and policy strategies that are consistent with the conceptions of economic justice that heterodox approaches value. Greater attention must be paid to the relationship between positive and negative economic rights, for instance. A second question concerns the kinds of policy measures that might generate equality of positive freedoms in ways that are widely taken to be ethically defensible—even perhaps by those who would do worse under such arrangements. And what challenges are posed to the equalization of positive freedom by the dramatic increases in international economic integration over the past several decades? What kinds of policy initiatives are now necessary at the multilateral level to promote global equality? These are difficult questions, to be sure, but providing compelling answers to them is vital to the success of heterodox research and political projects.

## Conclusion

This chapter has established that distinct approaches to economics value distinct notions of rights and freedom, which in turn generate distinct notions of justice, and these differences bear heavily on their respective assessments of economic policies and outcomes. Neoclassical thought entails a strong aversion toward value judgments, and this leaves it with little to say about economic justice. That said, there is implicit in the tradition a commitment to negative rights and negative freedom. Those who embrace this conception of freedom (such as economist Friedman and philosopher Nozick) are led to view any economic outcome as just, provided it arose through just means, where just means is defined as voluntary exchange, free of coercion. In this conception, then, free market outcomes are just because they arise from the exercise of people's rights. Even grossly unequal distributions of income are beyond reproach, provided they arose from processes (such as free exchange) that violated no one's rights.

In contrast, many heterodox traditions explicitly engage moral judgments, and many of these adopt a positive conception of rights and freedom. These approaches place emphasis on what a person can actually be or do, and many also view a just outcome as one that entails equality in positive freedom.

The case of bargaining between capital and labor examined in this chapter is just one example of what Marxists and other heterodox economists view as a general ethical problem in free market economies. Those who enjoy greatest substantive freedom are in positions to extend their own freedoms at the expense of those with least substantive freedom, because the enjoyment of substantive freedom facilitates greater bargaining power in market exchanges. Those with the greatest income, wealth, connections, and so forth will gain at the expense of those who lack these resources. This leads to the conclusion that the equality of negative rights that the free market enshrines may often deepen inequality in the positive freedoms that people enjoy.

## References and Further Readings


Anderson, E. (1990). The ethical limitations of the market. *Economics and Philosophy, 6*(2), 179–205.

Berlin, I. (1958). *Two concepts of liberty.* Oxford, UK: Oxford University Press.

Bowles, G., & Gintis, H. (1990). Contested exchange: New microfoundations of the political economy of capitalism. *Politics and Society, 18*(2), 165–222.

Burczak, T. (2001). Ellerman's labor theory of property and the injustice of capitalist exploitation. *Review of Social Economy, 59*(2), 161–183.

Commons, J. R. (1924). *Legal foundations of capitalism.* New York: Macmillan.

DeMartino, G. (2000). *Global economy, global justice: Theoretical objections and policy alternatives to neoliberalism.* London: Routledge.

DeMartino, G. (2003). Realizing class justice. *Rethinking Marxism, 15*(1), 1–31.

Dugger, W. (1989). Radical institutionalism: Basic concepts. In W. Dugger (Ed.), *Radical institutionalism* (pp. 1–20). New York: Greenwood Press.

Dugger, W., & Sherman, H. J. (1994). Comparison of Marxism and institutionalism. *Journal of Economic Issues, 28*(1), 101–127.

Ellerman, D. (1992). *Property and contract in economics.* Cambridge, MA: Blackwell.

Friedman, M. (1953). The methodology of positive economics. In M. Friedman (Ed.), *Essays in positive economics* (pp. 3–43). Chicago: University of Chicago Press.

Friedman, M. (1962). *Capitalism and freedom.* Chicago: University of Chicago Press.

Friedman, M., & Friedman, R. (1990). *Free to choose: A personal statement.* San Diego, CA: Harcourt Brace Jovanovich.

Geras, N. (1985, March/April). The controversy about Marx and justice. *New Left Review, 150,* 47–85.

Geras, N. (1992, September/October). Bringing Marx to justice: An addendum and rejoinder. *New Left Review, 195,* 37–69.

Locke, J. (1690). *Two treatises of government* (2nd ed., P. Laslett, Ed.). Cambridge, MA: Cambridge University Press.

Lukes, S. (1987). *Marxism and morality.* Oxford, UK: Oxford University Press.

Lutz, M. A. (1999). *Economics for the common good: Two centuries of economic thought in the humanistic tradition.* London: Routledge.

MacPherson, C. B. (1962). *The political theory of possessive individualism.* Oxford, UK: Oxford University Press.

Marx, K. (1938). *Critique of the Gotha Program.* New York: International Publishers.

Nozick, R. (1974). *Anarchy, state and utopia.* New York: Basic Books.

Nussbaum, M. (1992). Human functioning and social justice: In defense of Aristotelian essentialism. *Political Theory, 20*(2), 202–246.

Ramstad, Y. (1987). Free trade versus fair trade: Import barriers as a problem of reasonable value. *Journal of Economic Issues, 21*(1), 5–32.

Rawls, J. (1971). *A theory of justice.* Cambridge, MA: Harvard University Press.

Resnick, S., & Wolff, R. D. (1987). *Economics—Marxian versus neoclassical.* Baltimore: Johns Hopkins University Press.

Roemer, J. E. (1988). *Free to lose.* Cambridge, MA: Harvard University Press.

Sen, A. K. (1987). *On ethics and economics.* London: Blackwell.

Sen, A. K. (1992). *Inequality reexamined.* Cambridge, MA: Harvard University Press.

Tool, M. R. (1979). *The discretionary economy.* Boulder, CO: Westview Press.

Walzer, M. (1973). In defense of equality. *Dissent, 20*(4), 399–408.

Walzer, M. (1983). *Spheres of justice.* New York: Basic Books.

# 52

# Sports Economics

Aju Fenn
*Colorado College*

Sports economics is arguably the most popular undergraduate elective in the collegiate economics curriculum today. Of the top 50 liberal arts colleges surveyed, 61% of the respondents offered an elective in sports economics. Details of the survey are available upon request. The rankings are based on the *U.S. News & World Report* rankings ("Liberal Arts Rankings," 2009). Of the top 50 economics departments at research universities surveyed, 43% of the respondents offered a course in sports economics (Dusansky & Vernon, 1998). The *Journal of Sports Economics* has added special issues to keep up with the flow of scholarship, and more than one textbook has emerged on the subject. What is sports economics? Why might a topic such as sports stimulate pedagogical and scholarly interest among academics and students who usually pursue more serious subdisciplines, such as econometrics and mathematical economics?

Sports economics is the study of the allocation of scarce resources among competing desires in the context of sports. Although this definition is not very different from the definition of economics itself, it does reveal how the subdiscipline of sports economics was born. It also reveals the breadth of the field. Anything that carries the title of sports, from professional football to a lumberjack competition on ESPN, has the potential to stimulate a paper in this subdiscipline. Established economic scholars from other disciplines brought their standard tool kits to bear on professional and amateur sports data sets. At first, they studied baseball extensively because it was a sport that they had played and followed. Many of these scholars were tenured at prestigious research universities but undertook these projects because the sports industry was fun to study. At first, there were no textbooks for sports economics classes, and syllabi looked suspiciously like a list of applied microeconomics topics. Labor economists studied everything from wage discrimination to managerial efficiency. Industrial organization scholars promptly investigated the concept of market power on the field in wins and off the field in dollars, and environmental economists used contingent valuation methodology to determine the value of a sports team to a city or region. This chapter discusses the various strands of the literature in detail.

There are also other factors that contributed to the emergence of sports economics. Department chairs discovered that offering an elective in sports economics with a principles prerequisite was a good way to boost departmental enrollments. Other faculty members discovered that students who would normally run screaming at the mention of regression would sit patiently through the explanation of a multiple regression model that explained the determinants of competitive balance in the National Football League (NFL). With the advent of the Internet, sports data sets became readily available. Other students wanted to conduct applied econometric research in the context of sports with the aid of user-friendly econometrics software packages. At the Western Economics meetings (the formal partner for meetings of the North American Association of Sports Economists), the 8:00 a.m. sports economics sessions seemed to be drawing a crowd. Once in a while, participants would catch the guilty look of a macroeconomist who was having a little too much fun at that particular session. Economists are onto something here. Economics has not looked this appealing since the IS-LM model.

The rest of this chapter proceeds as follows. The various strands of the sports economics literature are discussed in each of the subsequent sections, as they apply to sports economics theory; the later sections discuss applied work,

policy implications, and directions for future research. The major strands of the literature include the theory of the firm, the theory of the consumer, economic impact studies, discrimination in sports, and collegiate sports. In each instance, the chapter strives to point out the classic readings and some more current work in the area. A brief section covers some of the many sports economists who have been instrumental in the birth and progression of the discipline. The chapter concludes with some discussion about the future directions of the field. First, the chapter turns to a loose classification of topics in sports economics.

## A Loose Classification of Topics in Sports Economics

Any classification of sports economics along traditional *Journal of Economic Literature* lines will be challenged by the differences between professional sports, amateur sports, and recreational sports. To be clear, professional sports are those where the contestants are paid for participation, amateur athletes are not directly paid for their participation by the contest organizers, and recreational sports are those undertaken for pure consumption value of participation and perhaps for the health benefits from exercise. Rather than attempt to construct an airtight classification that will be rendered obsolete by the evolving nature of the discipline, this chapter attempts to classify the literature based on the underlying big ideas in economics, such as the theory of the firm, the theory of the consumer, public policy regarding funding of stadiums, competitive balance, and discrimination in sports. In sports economics, as in economics, new branches often emerge from the cross-pollination of these big ideas. Any discussion of the classification of the literature on sports economics should begin with the original paper that brought economic ideas to bear on the sports industry. It is to that paper that this chapter now turns.

### Origins of Sports Economics

Sports economists agree that the original work in the field was Simon Rottenberg's paper on the labor market for baseball players (Rottenberg, 1956). In this paper, Rottenberg describes the existing rules of the baseball industry and their implications for competition. He discusses concepts such as the reserve clause, territorial rights, the drafting of minor league players into the majors, competitive balance, and market size. He also introduces the notion of a production function as it applies to sports, where the number of games (weighted by revenue) put on by a team is a function of its players and all other inputs, such as competing players, managers of both teams, transportation, and the ballpark, are considered a second factor of production. Rottenberg discusses the reserve clause—which allows the team to renew a player's contract at a wage set by the team, at not less than 75% of the current year's salary. He dismisses the argument that the reserve clause is desired to protect small-market teams (teams with smaller fan bases and lower revenues) from higher-paying large-market teams. Rottenberg cites the domination in the count of the number of pennants won by the Yankees and the St. Louis Browns as evidence of unequal player talent distribution. He also discusses territorial rights—the exclusive right to be the only Major League Baseball (MLB) team in a region—and the notion of competitive balance. Competitive balance refers to the ability of teams to have a roughly equal chance of winning a game. This topic is related to the uncertainty of output hypothesis, which maintains that there is greater fan interest when the outcome of who will win the event is fairly uncertain. Rottenberg discusses player drafts, territorial rights, and the reserve clause, which are all factors that impact the distribution of player talent and consequently impact the competitive balance in a league. Rottenberg then closes with his free market prescription for exciting and close baseball games. Many of the ideas raised by Rottenberg have grown into branches of sports economics today. For example, competitive balance has been an area of study in all of the major professional sports in the United States. Allen Sanderson and John Siegfried (2006) provide a 50th-anniversary perspective on Rottenberg's original work and the strength of his conclusions after 50 years of industry developments.

## The Theory of the Sports Firm

The key distinction between sports teams and rival firms in traditional economics models is as follows: Although sports teams seek to compete on the playing fields, they need their competitors to survive financially in order to put on a game or a match. This key point serves as the springboard for many of the departures from standard microeconomic models of the firm when applied to sports teams and leagues. Sports economics models of firms are based on three big ideas: (1) profit maximization behavior of sports teams and leagues, (2) market power of sports teams as both sellers of a unique product and almost exclusive employers of a highly skilled set of professional athletes, and (3) firms' decision-making process about hiring players and coaches. This chapter turns to each of these big ideas regarding the sports firm in as much detail as a single chapter will allow.

### Profit Maximizing Behavior

Several of the economic models treat a single team as the firm. Alternatively, some models deal with the league as a cartel or social planner, and each individual team is considered to be a member of this group. There is a healthy

literature on modeling both firms and leagues in the context of competitive balance. Papers in this area cover the measurement of competitive balance and the optimal policies to promote competitive balance in a league. Those studies are discussed in the section on competitive balance.

A simple model of static profit maximization of a sports firm is contained in the leading undergraduate textbook on sports economics. Michael Leeds and Peter von Allmen (2008) define profits as the difference between revenues and costs. Team revenues are broken down into gate revenues, broadcast revenues, and licensing and other stadium-related revenues, such as concessions and naming rights. Costs are usually dominated by players' salaries in the major leagues. Other costs include items such as travel, marketing, and administrative costs, including league costs.

There are more dynamic and complex models available as well. Most of these models deal with profit maximization in the context of multiple teams and the optimal policy for promoting competitive balance in a league. For example, Mohamed El-Hodiri and James Quirk (1971) develop a model of profit maximization for a sports team over time. They conclude that profit maximization of a given team and the promotion of competitive balance across a league are incompatible. Donald Alexander and William Kern (2004) explore some of the more dynamic elements that impact the franchise value of a team, such as team relocation, the presence of a new facility, market size, and regional identity.

In the traditional theory of the firm, the firm chooses its output level to maximize profits by taking market price to be either given in a competitive setting or subject to the constraint of the market demand curve if the firm is a monopolist. There is little debate over what exactly constitutes a unit of output. If a firm is making and selling jeans, for example, output is measured in terms of the number of pairs of jeans that are produced and sold. Another firm theory concept is the choice of product quality that a firm must consider. In sports economics, one has to consider the notion of output carefully. Is it just the number of games played by a team, or is it the winning percentage of the team that matters for profits? Is it merely winning or closeness of the outcome that boosts revenues? Thus, in sports, the concepts of quantity and quality are inextricably linked. In most other markets, one can segment the market by quality levels and then examine the output decision of the firm at a given choice of quality level. In sports, the number of games is not a team choice variable. It is set by the league. The quality of a team that a franchise chooses to field, however, is a choice variable. So it all comes down to two questions. Does winning matter (and at what cost)? What about the owner's utility maximization problem?

Andrew Zimbalist (2003) provides an overview of the profit maximization debate. He concludes that owners' objectives vary among leagues and that further research is needed. Alternative hypotheses of ownership do not preclude profit maximization but also include the owners' utility as part of the objective function. Such utility may come from being seen as a mover and shaker in the big city, like Jerry Jones, the owner of the Dallas Cowboys, or simply one who gets to call the shots, like Al Davis, the owner of the Oakland Raiders. D. G. Ferguson, Kenneth Stewart, J. C. Jones, and Andre Le Dressay (1991) examine the premise of profit maximization and find support for profit-maximizing behavior. Ferguson et al. assume that profit maximization is synonymous with revenue maximization.

The main challenge in doing empirical work in this area is to come up with good data on the cost side of the equation. Team revenues based on attendance and average ticket prices can be easily calculated. Broadcast revenues and revenue-sharing agreements are made public as well. Player salaries are usually available, but bonuses, travel costs, and other general and administrative costs may not be as easily available for teams. How much does it cost a team to put on a single game? The lack of precise data on this subject can make profit maximization a tough hypothesis to test. However, for the interested reader, a simple win maximization model that should be accessible to advanced undergraduates is discussed in Stefan Kesenne (2006).

Sometimes, static profit maximization may not be validated by the data because of a combination of the owners' desires to win and their goals to increase the long-term value of the franchise. Often, owners will try to break even each year while investing the profits back into player talent, coaching staff, or facilities. In the minor leagues such as AAA baseball, arena football, or lacrosse, the audience goes for the experience, and the weather may make a bigger impact on the attendance than the earned run average of the starting pitcher.

## Market Power in Sports

Another characteristic of sports teams is that they usually possess some degree of market power in both the product and input markets. A team is usually the only seller of a particular professional sports product in the output market (monopoly) and often the only employer of professional athletes in that sport in that particular city (monopsony). Recent empirical evidence by Stacey Brook and Aju Fenn (2008) seems to suggest that NFL teams do possess market power over their consumers. When college football athletes look at professional football as a career, there are a limited number of leagues or employers. Such an employment scenario with a single employer meets the classic definition of monopsony.

Lawrence Kahn (2000) provides an overview of the studies that discuss monopsony and other labor market issues in sports. Harmon Gallant and Paul Staudohar (2003) examine how antitrust law in the United States has influenced the evolution of professional sports leagues. Stephen Ross (2003) considers 10 important antitrust decisions where

courts have ruled against sports leagues and examines whether these decisions were in the best interest of the public from an economist's point of view. In general, economists argue that a monopolist will charge a higher than competitive price and will appropriate a portion of the consumer's surplus. Should the government then disband all operating professional leagues? What could possibly justify a legal monopoly in sports? Kahn (2003) argues that expansion to too many teams would lower player quality and that the optimum structure lies somewhere between a monopoly and a competitive solution. What complicates the issue further is that for every sold-out stadium in the NFL, there are additional fans at home that get to watch the game on television. This issue returns when this chapter explores the consumer side of sports economics.

Sports firms are supposed to hold power over their employees—and the players as well, because professional sports employment opportunities are somewhat limited. One of the oldest ideas in this area is the so-called invariance principle, which is an application of the Coase Theorem to sports leagues. Professional athletes have a unique set of skills and are in a position to generate an economic rent as a result. The invariance principle in sports economics states that regardless of whether the player or the owner controls the rights to the player, the mobility of players between teams should be the same. In the case of free agency, players reap the benefits by playing for the highest bidder. In the case of the team owning the rights to the player, the team may sell the player to the highest bidder. The origins of this idea are attributed to Simon Rottenberg (Sanderson & Siegfried, 2006). John Vrooman (2009) finds that if owners are win maximizing in nature, then they will erode their own monopsony power and compete aggressively for player talent. He states that the fact that most professional payrolls are about 60% of the revenue generated is evidence of the erosion of monopsony power by team owners. This is a budding area of policy in sports economics, but practical measurement of monopoly and monopsony power may be challenging because the details of cost and salary data are often private. This chapter turns next to the choices facing the sports team regarding talent evaluation and coaching.

## Input Decisions of the Firm: Which Players and Coaches to Hire?

Because sports teams spend vast amounts of money on player payroll, it stands to reason that most aspects of the labor market are well documented. It is not the number of units of labor (players) or human capital (coaches and players) that is important. In most leagues, these numbers are set by league rules. What matters here is how league owners and unions agree to split total revenue among owners and players and the ability of the team to evaluate the talent of players and coaches.

Rodney Fort (2006) has excellent coverage of the history of player pay, the value of sports talent, and labor relations in professional sports. The traditional explanation is that players are paid their marginal revenue products. In sports, this is often defined as the product of their marginal contributions to each win multiplied by the revenue earned for the franchise by that win. There is some disagreement among scholars about whether players' marginal revenue products are equal across teams. Stefan Szymanski and Kesenne (2004) argue that the marginal revenue of a win will be larger for a large-market team than for a small-market team.

So do players get paid their marginal revenue products of wins? What does the research say? In most leagues, such as the NFL, the league is a monopsony buyer of player talent, and there are some limits to compensation, such as a salary caps or disincentives such as a luxury tax. In a given league and within the limits of the salary cap, each team competes for the best players, thereby bidding up the wage for a player. In addition, players are often organized into unions so that overall negotiations between players unions and the team owners represent a bilateral monopoly (where a single buyer of talent, the team, faces a single seller of talent, the players union). In a bilateral monopoly, the equilibrium often comes down to the bargaining power of the two sides. If owners prevail in these negotiations, then through the use of salary caps or other restraints on compensation, players' wages are restricted to levels below their marginal revenue products. On the other hand, if players unions prevail in negotiations, then the aggregate share of total revenue that goes to the players becomes larger. Given these nuances, in sports it makes sense to review the evidence whether players are actually paid the marginal revenue product of wins.

Scully (1974) is among one of the first to study the issue of pay and performance in MLB. He finds that in the 1970s, baseball players were exploited by teams under the reserve clause, which prohibited players from seeking competitive employment with other teams. He finds that baseball players' average salaries over the length of their careers were only about 11% of their gross and 20% of their net marginal revenue products. Since then, many papers have been written about the determinants of wages and the impact of the type of contract (length, time in the contract, etc.) on performance. Alternatively, a more recent study by Vrooman (2009) finds that in most North American professional major leagues, players share about 60% of the revenues earned. Vrooman reports that all leagues except MLB have imposed salary caps just below 60% of league revenue. In short, with the advent of free agency, the balance of power between players and owners has evolved, and player salaries have risen.

What, then, are the determinants of an individual player's salary? Factors such as the performance of a player on and off the field, race, the revenue of the team, the type of arbitration scheme, the contributions of teammates, and league salary caps influence the salary that a player receives (Barilla, 2002; Berri & Krautmann, 2006; Brown

& Jepsen, 2009; Idson & Kahane, 2000). Kahn (2000) argues that the presence of rival leagues leads to higher salaries for players in baseball. Whenever competing leagues merge, the result is stronger monopsony power of owners over players and a decline in players' salaries. In most major league sports in North America, players are drafted by a team, which gives the team exclusive rights to negotiate with that player. Within a given league, it is presumably the ability of a player to contribute to wins and boost attendance that determines his or her level of salary.

Todd Idson and Leo Kahane (2000) examine the team effects on player compensation in the National Hockey League (NHL). They find that players' wages depend on their individual contributions and the impact of their teammates' play on their productivity. Idson and Kahane (2004) have gone on to study the themes of discrimination, market power, and compensation in their subsequent papers on the National Basketball Association and the NHL. They have also investigated the impact of characteristics of workers, such as language skills, on the pay and performance of coworkers (Simmons, Kahane, & Longley, 2009). There is also a large literature on the determinants of coaches' salaries and the retention of coaches. The interested reader is directed to the books mentioned for further reading at the end of this chapter.

The ability to spot talent is important in leagues where teams are constrained by league rules in the amount of money that they can spend on payroll. Player talent evaluation is often touted as the reason why small-market baseball teams such as the Oakland As and the Minnesota Twins can compete with large-market teams such as the Dodgers. The Patriot's three Super Bowl wins under coach Bill Belichick in the age of the salary cap is often attributed to the ability of the organization to spot talent. Contrary to the popular misconception that coach Belichick was a film major, it turns out that he actually majored in economics. What characteristics make certain players successful in the big leagues? Both front office general managers and sports economists alike would love to know the answer to this question. David Berri (2008) has devoted considerable effort to measuring productivity on the basketball court and examines how success in college may or may not translate into success in the professional leagues. J. C. Bradbury (2007) is to baseball sabernomics (the analysis of baseball using economic principles and econometric tools) what Berri is to productivity in the NBA. This is a budding area of research.

In summarizing the theory of the firm, one has to return to the question of whether sports firms maximize profits. The preponderance of evidence indicates that the majority of professional major league sports franchises seek to operate in the black in the short run while maintaining the goal of increasing the value of the franchise. There is also a certain utility associated with owning a major league team. However, to most owners, this utility is more than a hobby because they strive to make their franchises financially viable and competitive on the playing field at the same time. This chapter turns next to the individuals that are responsible for the growth of sports into big business: the fanatic or the supporter, as they are known in Europe.

## The Theory of the Sports Consumer

Readers may find this area easier to follow because most are consumers, if not sports consumers, and are thus familiar with the reasons and the ways in which consumers enjoy sports. Consumers attend sporting contests in person, watch them on television, or participate in sports. This chapter sets the participation aspects aside for now because the vast majority of consumers are not professional athletes. In each of these markets, fans that attend games or television audiences, the consumers, may be further divided into groups based on their intensity of preferences. Some fans are die-hard fans and live and die with the fortunes of their teams. Other casual fans, while interested, do not suffer these same highs and lows. Finally, there are those that happen to attend a sporting event or watch a game on television because it is just another entertainment option that they happened to choose. Some or all of these fans may choose to buy sports apparel either to proclaim their loyalty to their teams or as a fashion statement. All of these consumers have one thing in common. They are all maximizing their own utility functions.

### Attendance at Sporting Events

Fans may choose to attend games or matches because they believe that the experience will enhance their utility, subject to their budget and time constraints. Do fans have more fun watching their teams win, lose, or win in close games? Which outcome yields the most fan attendance?

Sports economists claim that winning is a very important determinant of attendance (Davis, 2009; Welki & Zlatoper, 1994). Most fans prefer a close contest, with their teams winning in the end. There has been much work done on the uncertainty of outcome hypothesis (Knowles, Sherony, & Haupert, 1992). The age of the sporting facility also matters. Brand new stadiums and arenas tend to draw more fans. John Leadley and Zenon Zygmont (2006) find that increased attendance due to a new stadium lasts for about 5 years. The prevailing wisdom about superstar players is that they promote attendance through winning at home and sell out games on the road because everyone wants to see them play (Berri, Schmidt, & Brook, 2004). The demand for sporting events in North America is considered to be unresponsive to changes in ticket price (Coates & Humphreys, 2007). Are sports fans addicts? According to traditional studies on rational addiction, the past and expected future consumption of a good should be significant determinants of current consumption of that good. In other words, if a fan has attended the games for a

given team in the past and plans on attending games in the future, that will impact the fan's decision to attend games in the present. Fans are like addicts in that watching games provides excitement. Fans experience exhilaration when their teams score and feel down when their teams fall dramatically behind. Die-hard fans follow the fortunes of their teams throughout the off-season and miss the Sunday ritual of watching their favorite NFL teams. During the season, they plan on watching their teams every Sunday. It is in this sense that fans are likened to addicts (Becker & Murphy, 1998). Sports economists have begun to consider this question as well. Young Lee and Trenton Smith (2008) find evidence that Americans are rationally addicted to baseball while Koreans are not. The literature in this area is fairly thin because it is a recent development in the field.

### Sports on Television

Most major league sports are broadcast on television in North America. The NFL has a blackout rule to prevent a reduction in ticket sales. If a game is not sold out 72 hours before kickoff, then it is blacked out in the local television viewing area. It tends to be the case that winning teams seldom have to worry about this rule, and about 90% of the games are sold out. In Europe, however, televising a game sometimes depresses attendance (Allan & Roy, 2008). The NFL currently has contracts with the major television networks to broadcast games through 2011. The value of these contracts is about $20.4 billion. This aspect of television viewing tends to be in the category of a public good. If enough people attend a game, a viewer can sit at home and watch the game without paying for a ticket. In some senses, this makes U.S. local television broadcasts nonrival and nonexcludable. The category of pay-per-view (PPV) is a little different. If one wants to watch an NFL team that is not being carried on the local television stations, one may have to purchase a special package from a cable or satellite television provider. Similarly, one may also purchase the opportunity to view certain other sporting events, such as boxing matches or soccer matches, which are available only on PPV. Either way, it is big business, and teams are able to extend their audiences beyond the confines of their stadia. There is a large literature on the economics of broadcasting and how certain systems impact consumer welfare. The consensus view among sports economists is that leagues that tend to share national television revenue equally, like the NFL, have greater competitive balance and consequently a greater demand for their product. On the other hand, leagues that have revenue imbalances, like MLB, where large-market teams like Los Angeles and New York have larger broadcast revenues, have less competitive balance and consequently a lower demand for their product. However, there is not much published academic research on the determinants of television ratings for sports events in North America. One reason for this may be that the data are proprietary information, and researchers may have to purchase data sets from a media research organization. This may be a potential area of expansion for the academic literature.

### Sports Merchandise and Memorabilia

Consumers purchase sports jerseys, baseball hats, and other assorted items. Other consumers are collectors and purchase items such as baseball cards and other autographed memorabilia. There is a literature on baseball cards and the impact of a player's performance and race on sales. However, the impact of championships or star players on sports apparel has yet to be investigated.

When one brings sports firms and consumers together, one often gets into the arena of public policy. It is to this subject that this chapter next turns its attention.

## Sports and Public Policy

Even those who do not care about sports will probably have a few thoughts on the subject when asked whether they think their taxpayer dollars should be used to fund a new stadium for their local professional team. In the 1950s, most professional sports teams played in privately owned stadiums or arenas. Most professional football teams were the tenants of professional baseball teams and played football games around the baseball schedule. All professional hockey teams played in private arenas. Many professional basketball teams played in college arenas and played their games around the collegiate schedule. In the 1990s, U.S. cities spent $5,298 million on 57 new venues in the four major professional sports. The public's share averaged $218 million for each of these venues. This is approximately 66% of the cost (Depken, 2006). This section considers some of the economic arguments presented for and against public funding for stadiums. Sanderson (2000) provides an excellent overview of this debate.

Proponents of public funding argue their cases on the basis of indirect and direct benefits of the team to the area. Indirect benefits—or benefits not accruing to the team—include the multiplier effect of job creation in the area due to team- and stadium-related activities. Teams often claim that the new stadium will be an engine of economic growth and revitalization for an area. Games draw crowds, and those crowds need to eat, drink, and shop. The direct benefits of a new stadium are those that accrue directly to the team and their fans. Teams contend that with the revenue from a new stadium, they can afford better players and contend for a championship. They claim that a new stadium enhances civic pride from living in a major league city. Last but not least, a new stadium would keep the team in town, and fans that attend games would retain their entertainment values, as would the fans that watch the televised games at home.

Critics of public funding for stadiums argue that the multiplier effect is overstated because of the so-called substitution effect. The substitution effect occurs when fans substitute attendance at sports events for other entertainment options like a movie at their local mall. Thus, the economic impact in the stadium area comes at the cost of spending at other entertainment venues. Critics claim that the benefits to consumers are not large enough to justify the subsidies given to sports teams. The consumer surplus generated from attending games is not large enough to justify the expenditures required to construct new stadiums (Alexander, Kern, & Neill, 2000). Critics also claim that stadium moves not only increase revenues through higher prices and attendance but also lower costs through favorable rental agreements. Most rental agreements provide attendance-based rents. This shifts the risk to the landlord, which in this case is the taxpayer. Robert Baade, Robert Baumann, and Victor Matheson (2008) find that megaevents such as the Super Bowl have no statistically significant impact on taxable sales. Yet in the end, city after city builds stadium after stadium for professional sports teams. Why is this so? Fenn and John Crooker (2009) examine the willingness of Minnesotans to pay for a new Vikings stadium given the credible threat of team relocation. They find that on average, households are willing to pay approximately $530 toward a new Vikings stadium. These results were obtained from a representative urban and rural sample of 1,400 households in Minnesota. There are two plausible explanations for the stadium building boom. Either the civic pride aspects have been undervalued or stadium advocates have been politically more successful at outmaneuvering their critics. Past studies fail to find any statistically significant relationship of the impact of a stadium on the income in the standard metropolitan statistical area (Baade & Dye, 1988). Similarly, Brad Humphreys (1999) analyzes data from every U.S. city that had a professional football, basketball, or baseball franchise over the period from 1969 to 1994. He finds that, contrary to the claims of proponents of sport facility subsidies, the presence of a professional sports team or facility has no effect on the growth rate of local real income per capita, and it reduces the level of local real income per capita by a small but statistically significant amount. The problem with all the studies done on valuing civic pride is that they are surveys with no binding commitment on the part of the respondents to actually spend the money. There is room for a study that uses the experimental economics approach by giving participants a sum of money and a credible scenario of a relocation to see how much money they actually donate. Jesse Ventura, governor of Minnesota from 1999 to 2003, asked people to turn their tax rebates in to support funding a new stadium. The electorate greeted him with the usual response of bewildered amusement.

Having discussed the sports firm, the sports consumer, and how professional sports impacts public policy, this chapter turns to issues of competitive balance, discrimination, and collegiate sports.

## Competitive Balance in Sports

Competitive balance refers to a situation where teams have a more or less equal chance of winning a game. This does not necessarily mean that all teams in the league must be of the same talent level. The NFL, in fact, takes past success into account while making the schedule for the next year. They pit stronger teams against stronger opponents, and weaker teams get easier opponents. Why should one care about competitive balance? U.S. sports economists argue that competitive contests are what drive attendance. European sports economists are not as concerned with competitive balance. In the English Premier League, for example, teams at the top vie for championships, while teams at the bottom strive to avoid relegation. The supporters are regionally loyal to their teams. They would love to win, but winning is not all that matters. The competitive balance literature bifurcates into two major strands. The first branch deals with constructing indices to measure and observe competitive balance. The second deals with policy prescriptions to promote competitive balance.

Sanderson and Siegfried (2003) present a useful review of the different measures of competitive balance. Measures range from simple measures of dispersion, such as standard deviations of winning percentages around league means, team means over time or at a point in time, modifications of Gini indices, and the deviations of the Herfindahl-Hirschman Index. There is also a considerable literature on the use of these measures in North American major league sports.

The second branch of this literature deals with both the measurement and policy prescription for a more competitive league via practices such as revenue sharing and the reverse order draft. For example, Andrew Larsen, Fenn, and Erin Spenner (2006) find that the NFL did indeed become more competitive as a league after the institution of the salary cap and free agency in 1993. There are also theoretical models that examine how competitive balance in a league may be improved. Crooker and Fenn (2007) examine the state of competitive balance in MLB and provide a league transfer payment mechanism for improving parity and, consequently, league profits.

## Discrimination in Sports

Given that one can observe a player's performance and compensation with some degree of clarity, sports economists have used the data to test for racial and gender discrimination in sports, both in the professional and collegiate arenas. Sports economics textbooks such as

Leeds and von Allmen (2008) classify discrimination into employer discrimination, employee discrimination, consumer discrimination, gender discrimination, and positional discrimination. The interested reader is referred to their book for an extensive discussion of the theory and applied work on discrimination in sports economics. In general, most studies regress wages against performance statistics and race and gender variables to examine the role of race and gender. Other studies cover the values of baseball trading cards and the race of the player involved. The big idea in this area is that discrimination did exist in the past, both in terms of salaries and consumer preferences. For example, Mark Kanazawa and Jonas Funk (2001) find that predominantly white cities enjoy watching white players play for their hometown NBA team. However, both employer and consumer discrimination has been decreasing in recent years. Title IX and its influence on gender balance in collegiate sports is another big idea that dominates the literature on discrimination and college sports.

## Collegiate Sports

The big elephant in the room is the unpaid professional: the college athlete. Collegiate athletic directors claim that college athletes for big-time programs bring in revenues that are redistributed to other programs in the athletic department and sometimes even to the rest of the school. The financial details of National Collegiate Athletic Association (NCAA) Division I programs are available in a report from the NCAA (2008).

Title IX and its impact on collegiate athletics have dominated the literature recently. Sports economists have also become fascinated with the question of a national playoff in NCAA football. Zimbalist (2001b) has an excellent book on the subject of college sports that covers the relevant issues. This topic is a chapter unto itself and goes well beyond the scope of an overview of sports economics.

## Conclusion

The world of sports economics is constantly evolving. Some topics, such as the Olympics, bowling, golf, NASCAR, professional bass fishing, and distance running have not been covered in this chapter because of space limitations There is, however, a literature on each of these sports. Among the issues that present excellent opportunities for new research are the connections between gambling and sports. In particular, are there aspects to watching sports that are addictive? Another area of interest is the connection between sports and the joint consumption or depreciation of an individual's health stock: Some fans may choose to drink beer while watching sports, and others may be motivated to participate in adult recreational sports leagues after watching an exciting contest. The economics of youth sports and adult recreational sports has been largely unexplored; this will also be an area of interest in the years to come. Older topics that pertain to the impact of institutions and rules on sports, such as free agency and drug testing, will be examined time and again as the institutions and rules evolve. The economic impact of a sports team on a region will remain a constant topic of research as long as teams seek public funding. Studies and statistics on unlocking the keys to winning will also be around for quite some time. If a sport is televised and people watch it, sooner or later a sports economist will analyze it. Be prepared for the first study on competitive balance in the World's Strongest Man competition; it could happen.

*Author's Note:* This chapter is dedicated to the memory of Larry Hadley.

## References and Further Readings


Alexander, D. L., & Kern, W. (2004). The economic determinants of professional sports franchise values. *Journal of Sports Economics, 5*(1), 51–66.

Alexander, D. L., Kern, W., & Neill, J. (2000). Valuing the consumption benefits from professional sports franchises. *Journal of Urban Economics, 48*(2), 321–337.

Allan, G., & Roy, G. (2008). Does television crowd out spectators? New evidence from the Scottish Premier League. *Journal of Sports Economics, 9*(6), 592–605.

Andreff, W., & Szymanski, S. (Eds.). (2006). *Handbook on the economics of sport.* Northampton, MA: Edward Elgar.

Baade, R. A., Baumann, R., & Matheson, V. A. (2008). Selling the game: Estimating the economic impact of professional sports through taxable sales. *Southern Economic Journal, 74*(3), 794–810.

Baade, R., & Dye, R. (1988). An analysis of the economic rationale for public subsidization of sports stadiums. *Annals of Regional Science, 22*(2), 37–47.

Barilla, A. G. (2002). *An analysis of wage differences in Major League Baseball, 1985–95.* Unpublished doctoral dissertation, Kansas State University.

Becker, G., & Murphy, K. (1998). A theory of rational addiction. In K. Lancaster (Ed.), *Consumer theory* (Vol. 100, pp. 581–606). Northampton, MA: Edward Elgar.

Berri, D. J. (2008). A simple measure of worker productivity in the National Basketball Association. In B. Humphreys & D. Howard (Eds.), *The business of sport* (pp. 1–40). Westport, CT: Praeger.

Berri, D. J., & Krautmann, A. C. (2006). Shirking on the court: Testing for the incentive effects of guaranteed pay. *Economic Inquiry, 44*(3), 536–546.

Berri, D. J., Schmidt, M., & Brook, S. (2004). Stars at the gate: The impact of star power on NBA gate revenues. *Journal of Sports Economics, 5*(1), 33–50.

Bradbury, J. C. (2007). *The baseball economist: The real game exposed.* New York: Dutton.

Brandes, L., Franck, E., & Nuesch, S. (2008). Local heroes and superstars: An empirical analysis of star attraction in German soccer. *Journal of Sports Economics, 9*(3), 266–286.

Brook, S. L., & Fenn, A. J. (2008). Market power in the National Football League. *International Journal of Sport Finance, 3*(4), 239–244.

Brown, K. H., & Jepsen, L. K. (2009). The impact of team revenues on MLB salaries. *Journal of Sports Economics, 10*(2), 192–203.

Coates, D., & Humphreys, B. R. (2007). Ticket prices, concessions and attendance at professional sporting events. *International Journal of Sport Finance, 2*(3), 161–170.

Crooker, J. R., & Fenn, A. J. (2007). Sports leagues and parity: When league parity generates fan enthusiasm. *Journal of Sports Economics, 8*(2), 139–164.

Davis, M. C. (2009). Analyzing the relationship between team success and MLB attendance with GARCH effects. *Journal of Sports Economics, 10*(1), 44–58.

Depken, C. A. (2006). The impact of new stadiums on professional baseball team finances. *Public Finance and Management, 6*(3), 436–474.

Dusansky, R., & Vernon, C. J. (1998). Rankings of U.S. economics departments. *Journal of Economic Perspectives, 12*(1), 157–170.

El-Hodiri, M., & Quirk, J. (1971). An economic model of a professional sports league. *Journal of Political Economy, 79*(6), 1302–1319.

Fenn, A. J., & Crooker, J. R. (2009). Estimating local welfare generated by an NFL team under credible threat of relocation. *Southern Economic Journal, 76*(1), 198–223.

Ferguson, D. G., Stewart, K. G., Jones, J. C., & Le Dressay, A. (1991). The pricing of sports events: Do teams maximize profit. *Journal of Industrial Economics, 39*(3), 297–310.

Fort, R. (2006). *Sports economics.* Upper Saddle River, NJ: Prentice Hall.

Gallant, H., & Staudohar, P. D. (2003). Antitrust law and public policy alternatives for professional sports leagues. *Labor Law Journal, 54*(3), 166–179.

Gerrard, B. (2007). *Economics of association football.* Northampton, MA: Edward Elgar.

Humphreys, B. (1999). The growth effects of sport franchises, stadia and arenas. *Journal of Policy Analysis and Management, 18*(4), 601–624.

Humphreys, B. (2002). Alternative measures of competitive balance in sports leagues. *Journal of Sports Economics, 3*(2), 133–148.

Idson, T., & Kahane, L. (2000). Team effects on compensation: An application to salary determination in the National Hockey League. *Economic Inquiry, 39*(2), 345–357.

Idson, T., & Kahane, L. (2004). Teammate effects on pay in professional sports. *Applied Economic Letters, 11*(12), 731–733.

Kahn, L. M. (2000). The sports business as a labor market laboratory. *Journal of Economic Perspectives, 14*(3), 75–94.

Kahn, L. M. (2003). *Sports league expansion and economic efficiency: Monopoly can enhance consumer welfare* (CESifo Working Paper No. 1101). Munich, Germany: CESifo Group Munich.

Kanazawa, M. T., & Funk, J. P. (2001). Racial discrimination in professional basketball: Evidence from Nielsen ratings. *Economic Inquiry, 39*(4), 599–608.

Kesenne, S. (2006). The win maximization model reconsidered: Flexible talent supply and efficiency wages. *Journal of Sports Economics, 7*(4), 416–427.

Knowles, G., Sherony, K., & Haupert, M. (1992). The demand for Major League Baseball: A test of the uncertainty of outcome hypothesis. *American Economist, 36*(2), 72–80.

Larsen, A., Fenn, A. J., & Spenner, E. L. (2006). The impact of free agency and the salary cap on competitive balance in the National Football League. *Journal of Sports Economics, 7*(4), 374–390.

Leadley, J. C., & Zygmont, Z. X. (2006). When is the honeymoon over? National Hockey League attendance, 1970–2003. *Canadian Public Policy, 32*(2), 213–232.

Lee, Y. H., & Smith, T. G. (2008). Why are Americans addicted to baseball? An empirical analysis of fandom in Korea and the United States. *Contemporary Economic Policy, 26*(1), 32–48.

Leeds, M., & von Allmen, P. (2008). *The economics of sports* (3rd ed.). Boston: Addison-Wesley.

Liberal arts rankings. (2009). *U.S. News & World Report.* Available at http://colleges.usnews.rankingsandreviews.com/college/liberal-arts-search

NCAA Publications. (2008). *2004–2006 NCAA revenues and expenses of Division I intercollegiate athletics programs report.* Available at http://www.ncaapublications.com/ProductsDetailView.aspx?sku=RE2008

Rosentraub, M. S., & Swindell, D. (2002). Negotiating games: Cities, sports and the winner's curse. *Journal of Sport Management, 16*(1), 18–35.

Ross, S. F. (2003). Antitrust, professional sports, and the public interest. *Journal of Sports Economics, 4*(4), 318–331.

Rottenberg, S. (1956). The baseball players' labor market. *Journal of Political Economy, 64,* 242.

Sanderson, A. R. (2000). In defense of new sports stadiums, ballparks and arenas. *Marquette Sports Law Journal, 10*(2), 173–192.

Sanderson, A. R., & Siegfried, J. J. (2003). Thinking about competitive balance. *Journal of Sports Economics, 4*(4), 255–279.

Sanderson, A. R., & Siegfried, J. J. (2006). Simon Rottenberg and baseball, then and now: A fiftieth anniversary retrospective. *Journal of Political Economy, 114*(3), 594–605.

Scully, G. (1974). Pay and performance in major league baseball. *American Economic Review, 64,* 915–930.

Shmanske, S. (2004). *Golfonomics.* River Edge, NJ: World Scientific.

Simmons, R., Kahane, L., & Longley, N. (2009). *The effects of coworker heterogeneity on firm-level output: Assessing the impacts of cultural and language diversity in the National Hockey League* (Working Paper No. 005934). Lancaster, UK: Lancaster University Management School, Economics Department.

Szymanski, S., & Kesenne, S. (2004). Competitive balance and gate revenue sharing in team sports. *Journal of Industrial Economics, 52*(1), 165–177.

Vrooman, J. (2009). Theory of the perfect game: Competitive balance in monopoly sports leagues. *Review of Industrial Organization, 34*(1), 5–44.

Welki, A. M., & Zlatoper, T. J. (1994). U.S. professional football: The demand for game-day attendance in 1991. *Managerial and Decision Economics, 15*(5), 489–495.

Zimbalist, A. (Ed.). (2001a). *The economics of sport.* Northampton, MA: Edward Elgar.

Zimbalist, A. (2001b). *Unpaid professionals: Commercialism and conflict in big-time college sports.* Princeton, NJ: Princeton University Press.

Zimbalist, A. (2003). Sport as business. *Oxford Review of Economic Policy, 19*(4), 503–511.

# 53

# Earnings of Professional Athletes

Lee H. Igel and Robert A. Boland
*New York University*

One of the greatest challenges facing professional sports has been the rapid increase in earning power of professional athletes during the past quarter century. This challenge is likely to dominate the sports business landscape in the coming decades, especially as salaries and endorsements that have reached averages in the millions of dollars encounter an increasingly turbulent, complex, and transnational economy. But now, more than ever before, the income potential of professional athletes has significant implications for the relationship among sports, business, and society.

As professional sports became more organized during the twentieth century, professional athletes in all developed countries were increasingly paid several times the average worker's salary. But even the relatively high-paying jobs that persisted through the first 70 or so years of the century did not parallel the rise, magnitude, and capacity of the professional athlete's earning power during the latter part of the century, especially in the professional baseball, football, basketball, and hockey leagues in the United States. The rise of the professional athlete paralleled that of the blue-collar worker, albeit with much better pay. That is, both professional athletes and blue-collar workers carved out a livelihood based on a cash wage with few fringe benefits until unionization increased their capacities to earn higher salaries and gain benefits. This increased both their social standing and their political power. But just as dramatically as the status of the blue-collar worker has fallen during the past 30-plus years, the status of professional athletes has experienced growth.

## The Evolution of Athletes' Salaries

To fully perceive the economics of modern athletes' earning power, it is helpful to bear in mind the rapid transformations that have occurred as sports shifted from pastime to business. There is perhaps no better starting point for such explanation than in the evolution of athletes' salaries, which have traditionally served as the greatest portion of their incomes.

Talk of athletes' salaries has often been considered in the context of team sports. Yet whether athletes play team or individual sports, professional athletes by definition receive pay for their performance in athletic competition. One of the early forms of paying individuals for athletic performance was established with the emergence of prizefighting in the late nineteenth and early twentieth centuries. In particular, the influx of European émigrés to the United States during this period of time, combined with the emancipation of slaves in the post–Civil War era, resulted in the ascension of social minorities through participation in sport. While white Irish Americans were the first to gain mass popularity—and to be paid accordingly—as prizefighters, Jack Johnson, a black man, had become heavyweight champion by the end of the first decade of the twentieth century. Johnson's ability to come out on the winning side of fight after fight elevated him to the status of pop culture icon, which brought with it all the benefits and trappings of such a position.

However, much of the standard for today's salary structures across the sports landscape has its modern

roots in the advent of professional baseball in the United States. During the late nineteenth century and for more than half of the twentieth century, baseball was the most popular professional team sport, and Babe Ruth was baseball's—and, arguably, the country's—most popular and dominant figure. Ruth, who played for the New York Yankees, was among a few stars whose presence on a team had such an impact on the press and at the turnstiles that he had the power to command an exceptionally high salary. And as one well-known anecdote relates, when reporters asked Ruth, whose contract at one point was valued at $80,000 per year, why he should be paid more than President Herbert Hoover, Ruth reportedly replied, "Why not? I had a better year than he did."

Nevertheless, the contracts of such stars were outliers. The average salary for a professional baseball player at the time was approximately $7,000, and although this figure was upward of five times that of the average working family, the disparity within the sport tells of how little impact stars had on the salary demands of other players on their teams. It also hints at the extent to which most players were bound to the whims of ownership.

As in almost any organization, payroll allocation has historically been one of the primary influences on sports franchises' decisions when negotiating player contracts. In fact, the infamous Black Sox scandal, in which eight members of the 1919 Chicago White Sox conspired with gamblers to intentionally lose that year's World Series, is generally considered to be the players' reaction to having felt underpaid by team owner Charles Comiskey. Yet the players' earnings were constrained not only by the actions of a penny-pinching owner. They were also severely limited by the existence of the reserve clause.

## The Reserve Clause and Free Agency

The reserve clause was one of the longest standing provisions in player contracts of both the major and minor leagues. It bound a player to a single team, even if he signed contracts on an annual basis. The term of reserve, therefore, effectively restricted a player from changing teams unless ownership granted his unconditional release from the team.

From the late 1800s until the 1960s, when the amateur draft was instituted, the only conceivable license that players had over their careers was the freedom to negotiate as an amateur with any team willing to sign them to a professional contract. Once the contract was signed, it was at the team's discretion to trade, sell, reassign, or release the player. The only alternative leverage that players held at contract time was to hold out—that is, refuse to play unless their preferred conditions were met.

One of the earliest serious studies on the reserve clause and its implications was conducted by Simon Rottenberg in a 1956 article, "The Baseball Players' Labor Market." Rottenberg concluded that the reserve clause inevitably transferred wealth from the players to the owners. He also determined that the best players tended to play for teams in the largest markets because these teams were in the optimal position to exploit the players' talent for the benefit of attracting fans to the ballpark.

But only a couple of years prior to the publication of Rottenberg's (1956) study, the Major League Baseball Players Association (MLBPA) had been established. As it functioned early on, the MLBPA was largely unassuming, to the point of being ineffectual until 1966 when the players hired Marvin Miller, a former negotiator for American steel workers, as the head of their union. The appointment of Miller would forever change the standards of player compensation—first in baseball and then across all professional sports.

On behalf of the players, Miller began to press for increases in such contract provisions as the minimum salary and pension contributions by owners. In so doing, he was progressively beginning to disrupt the basic assumptions of player contracts and, by extension, the relationship between ownership and players. Not long thereafter, things reached a tipping point when Miller challenged the legitimacy of the reserve clause.

In 1970, Curt Flood was a star player for the St. Louis Cardinals who had been traded to the Philadelphia Phillies. Flood, however, did not want to move from St. Louis and informed the Cardinals, Phillies, and the baseball commissioner's office of his intention to stay put and play out the remainder of his contract in St. Louis. Commissioner Bowie Kuhn ruled that such action was not within Flood's rights as a player and ordered him to play for Philadelphia or not play at all. Flood chose the latter and sued Major League Baseball (MLB) for violation of U.S. antitrust laws.

The case of *Flood v. Kuhn* (1972) eventually reached the U.S. Supreme Court, which ultimately sided with MLB. The Supreme Court cited, if somewhat dubiously, MLB's decades-old exemption from antitrust law. But losing in court did little to deter Miller and the players he represented. They instead took the owners head-on in a series of labor negotiations.

By 1972, a labor impasse boiled over when team owners refused to bargain with the players union on salary and pension issues. Now that they were firmly organized and unified by Miller, the players responded with the first leaguewide strike in American professional sports history. Only after nearly 100 games were lost to the strike did the owners finally concede to the players' demands. (The players found the labor stoppage so successful a tactic that they used it again in 1981, 1985, and 1994; the owners took a similar tack in 1976 and 1989, when they locked out the players during other labor disputes.)

As the players gained increasingly equal leverage in negotiations with owners during the 1970s, they pushed for a growing number of concessions. Of these, none was perhaps as influential to the earning power of professional athletes as the advent of free agency.

In 1974, Jim "Catfish" Hunter, a pitcher for the Oakland Athletics, became the first player to qualify for free agency. Hunter and team owner Charles Finley negotiated a contract that included a clause that required Finley to make a payment into an annuity for Hunter on a certain date. When Finley missed the date and subsequently attempted to pay Hunter directly rather than honor the clause, Hunter and Miller filed a complaint charging that the contract was null and void because Finley had broken the terms of the contract.

The case was sent to an arbitrator, who sided with Hunter. The voided contract made Hunter a free agent, which created a bidding war for his services. When Hunter signed a contract, he did so with the New York Yankees for a guaranteed salary that was precedent setting in both size and duration: $750,000 per year for 5 years. Even the immediate implications of the deal were far reaching: Hunter not only became the highest paid player in baseball history, but he also was one of the first players to receive anything more than a 1-year contract—and a guaranteed one at that. More important, the deal effectively established free agency across all of baseball.

As free agency took shape in the mid-1970s, the concept of the reserve clause was undergoing its final days of existence. In 1975, pitchers Andy Messersmith and Dave McNally played under the terms of the reserve clause. But when it came time for them to sign their contracts, they, with Miller's encouragement, argued that the reserve clause could not be applied if no contract was signed. Their case went before arbitrator Peter Seitz, who struck down the reserve clause and thereby resolved that the players could become free agents and sell their services to the highest bidder.

With the constraints of the reserve clause compromised and the rise of free agency, it was but a matter of time before these transformations would recast the expectation for player contract lengths and values and reconstitute the leitmotiv of the individual over the team.

## On the Implications of Free Agency

The players' gaining the right to freely offer their services to any team (on expiration of contract) had an astonishing impact on salaries. In 1975, the minimum salary in MLB was $16,000, and the average salary was $44,676; by 1980, the salaries had jumped to $30,000 and $143,756, respectively; within 10 years of the start of free agency, the salaries had risen to $60,000 and $371,571, respectively; and by the 20th anniversary, they had risen to $109,000 and $1,110,766, respectively. And while minimum and average salaries continued to rise year after year, nothing did more to demonstrate the extent of free agency than the 10-year, $252 million contract that Alex Rodriguez signed with the Texas Rangers in 2000.

The team's bid, which exceeded the next highest offer by approximately $100 million, personifies the theory that free agency is an auction market for athletes. According to work popularized by Richard Thaler (1992), this particular type of auction is a common value auction, in which the item being auctioned is more or less of equal value to all bidders, though bidders do not know the market value of the item when placing their bids. To place a bid, however, each bidder must have independently and expertly estimated the value of the item prior to bidding. When the auction is complete, the winner of the auction almost always is the one who provided the highest estimate. But as Thaler propounded, the winner of the auction ends up the loser because either the winning bid usually exceeds the true value of the item and the enterprise therefore loses money, or the true value of the item is less than the independently and expertly furnished estimate.

Despite most bidders' awareness of the "winner's curse," it fundamentally occurs because the thought processes of those doing the bidding—that is, ownerships—is irrational. The epitome of this reality is that the Rodriguez contract was eclipsed in 2007 when Rodriguez himself, who had by then been traded to the New York Yankees in 2004, signed a contract extension worth $275 million over 10 years. Most remarkable about the deal is not the numbers associated with it but that Yankees home games had been reaching sellout capacity and that every team Rodriguez played for had an improved win-loss record after he was no longer with them.

As such, the "winner's curse" is, in this context, of enormous benefit to the players. Because free agent salaries in professional sports are set from the top down, Hall of Fame–caliber players inevitably command the highest salaries. But because the free agent market persists on a scarcity of talent available in a given year, even second-level talent can expect to be compensated at a disproportionately higher amount, relative to their actual value of performance.

Although professional baseball was the forerunner of almost every significant policy for the movement of players between teams, the labor market for all athletes in every professional league is, fundamentally, a result of bargaining between owners and players. Over time, both owners and players have tended to agree that free agency is a means through which to provide the greatest perceived individual economic benefit. For owners, the costs associated with procuring and developing young talent has become so high as to insist that their best protection is through policy that requires a newly professional player be bound to his original team for a short period of time; in MLB, for instance, the period is 6 years. When this period of time lapses, the player qualifies for free agency and, under the rules of restricted free agency, can negotiate a new contract with any team, including his original one.

The conditions under which an athlete qualifies for restricted free agency vary among the major professional sports leagues. But the rules generally hold that although the athlete has freedom to negotiate with any team and agree on terms of a contract, the athlete's current team has an opportunity to match the terms of the deal before a contract is signed with the new team. Quite often, league rules specify that when a restricted free agent signs a deal with a new team, that team must compensate the athlete's prior team with an equitable number and level of picks in future drafts of amateur players. This framework is different from that of unrestricted free agency, under which the athlete has either been released outright from the team, not been offered a renewal upon expiration of contract, or not been selected in the amateur draft. Unrestricted free agents are, therefore, permitted to entertain and decide for themselves about contract offers from any team, an opportunity that can be especially lucrative for those who are the top performers in a league.

Yet even within the span of time between the beginning of a professional career and becoming eligible for free agency, and though it is in ways reminiscent of the reserve clause, the athlete is not entirely constrained by the whims of the owner. During this period, contracts are typically structured to provide players with annual salary increases, while annual minimum salary increases and salary cut percentages are mandated by the league, and as a result of the collective bargaining agreements between owners and players unions, players have the right to have their contract grievances with the owners ruled on by an independent arbitrator in the first couple or few years of the original deal. Given that these mechanisms essentially compare one player to other players in the league and their salaries and that the market value for players has increased over time, athletes stand to earn better than an "honest day's pay." In fact, they possess an increasing opportunity to create generational wealth for themselves and their families.

Since athletes have become more and more successful at increasing their salaries, a good many teams have in turn had to repeatedly relinquish key players. At the same time, in an attempt to remain competitive in both the game and the business, managements have discovered combinations of human intuition and statistical measurement that predict player performance. One result has been that a variety of teams in a number of leagues have begun to develop cost-containment strategies specific to their rosters.

Although Branch Rickey pioneered such systems during his years as an executive with the St. Louis Cardinals in the 1930s, perhaps the most popular of modern-day philosophies is that detailed in the book *Moneyball: The Art of Winning an Unfair Game* by Michael Lewis (2003). Lewis's treatment is an examination of Billy Beane, who as general manager of the Oakland Athletics MLB franchise, has used sabermetrics—mathematical examination of baseball-specific statistics—to make decisions about which amateur players to draft and which free agents to sign according to a notoriously low budget by league standards. With this system, which has the capacity to illuminate traits of the game that may not be immediately apparent, the team is theoretically able to obtain large numbers of undervalued, serviceable players at all positions. It is, in practice, a cost-containment strategy that helped Oakland compete with high-spending teams and qualify for the playoffs regularly in the 1990s and throughout the first decade of the 21st century, despite having the second lowest payroll in baseball.

By now, several other roster cost-containment theories have also been applied with success throughout professional sports. One of these involves the notion of free agency avoidance, in which teams essentially resist pursuit of high-priced free agents, including their own, and instead purposefully plan to replace those players with other talent already in the organization. Another theory involves teams extending substantial longer-term contracts, with submarket signing bonuses, to young players before they are eligible for free agency. This theory inherently provides an element of security to both the team and the player; they both effectively bet against the uncertainty of free agency.

The successful management of such human resources and salary allocation theories is evidenced in sports that operate both with and without salary caps. For example, in the National Football League (NFL), which does limit the amount of money a team can spend on total player salaries, teams such as the New England Patriots, Pittsburgh Steelers, and Philadelphia Eagles have implemented cost-containment theories and consistently appeared in the playoffs during the late 1990s and early 2000s. As a further example, Bill Belichick, head coach of the Patriots, and his staff have managed roster concerns by actively seeking players at positions less influenced by free agency and shying away from overvaluing and overpaying players at quarterback, wide receiver, and cornerback positions. Still, despite the relative success of these methods of controlling player costs, no team is immune from the reality of the rising costs of player contracts.

Although numerous factors contribute to the values of player contracts, the dollar amounts and conditions basically settle on that most fundamental of economics concepts: supply and demand. As referred to earlier, there is a limited supply of individuals who are able even in the first place to perform at the professional level. When a team plans to acquire an individual who is among the best performers in the sport, the decision makers in the organization must prepare to deal with the realities of smaller supply and higher demand.

According to application of economic theory, the supply curve's inelasticity in this relationship means the value of contracts will be determined by the demand curve. The increase of the demand curve expresses the rise in the individual player's value and, to be specific, the value of the individual player's contract. But this simple relationship

tells only part of how and why player contracts are valued as they are today. What is further said to explain current thinking about player contracts dates to the early days of modern economic theory: the neoclassicists' creation of marginal utility theory circa 1870.

Within marginal utility theory, which has taken to being called *microeconomics* since the establishment of Keynes's economic synthesis, there is the marginal revenue productivity (MRP) theory of labor and wages. Marginal revenue is the return obtained from the last unit sold and is a function of change in total revenue divided by change in quantity. As applied to labor and wages in general, it holds that workers are paid according to the value of their marginal revenues to the enterprise. In the context of player contracts, this means salary is based on and differentiated by performance, productivity, and output. More simply, players are paid based on whatever is considered to be their contributions to the enterprise.

Assuming that ownership is willing to pay a productive player for adding to the financial earnings of the franchise, what exactly is contained in the player's contribution? The answer has generally been to reduce MRP to the number of tickets sold to fans as a result of a player being a member of the team. But this is conceivably too simple an explanation. Given the amount of revenue streams flowing into the sports leagues and the business as a whole, it is increasingly difficult to ascertain a valid and reliable MRP for the professional athlete. This is in large part why those who have devoted a considerable amount of recent scholarship to understanding sports economics—certainly Andrew Zimbalist of Smith College and *The Wages of Wins* authors David Berri, Martin Schmidt, and Stacey Brook—have stated that salary structures are questionable and misguided and that there is by now a need to modify the way salaries are determined. At the heart of their arguments is the relationship among player salary, performance statistics, and which statistic matters most in determining player pay.

There is no doubt that the terms, considerations, and values of player contracts have experienced explosive growth since the advent of free agency. They have, in turn, brought significant change to sports and the sports business. Prior to free agency, professional athletes had generally been subject to the benevolent interests of owners; the typical athlete's acts of defiance amounted to some combination holding out for higher salary, overcoming the influence owners held over members of the press, and the occasional (though ultimately prohibited) acts of collusion in which owners tacitly agreed not to bid on each other's free agents. Today, professional athletes are arguably no longer exploited, inasmuch as athletes and owners have an increasingly balanced amount of control over the rules and courses of making money from sports. This is perhaps as much a result of successful labor negotiations as of rising exposure across the various means of communication—radio, television, newspapers, magazines, and the World Wide Web—that reach and influence the sports fan and the public at large.

## The Evolution of Athletes' Endorsements

In addition to the labor triumphs achieved by professional athletes in the last quarter of the twentieth century, there has been no greater or more powerful transformation than that of professional athletes basing their salary demands on the revenue streams afforded by television broadcasts. When television sets penetrated the media market and became an increasingly mainstream household item beginning in the 1950s, teams and leagues found a fresh medium by which to transmit their product to audiences, especially those that were outside the range of a particular radio signal. The broadcast of professional sports events was a natural fit for programming executives, who rapidly signed teams and leagues to lucrative contracts. But although these deals almost immediately increased the income and value of teams and leagues, only a very small minority foresaw the unintended consequence they would have on the incomes of professional athletes.

## The Impact of Television and Media on the Earnings of Professional Athletes

Because the boom in media rights deals made plenty of news, the public reporting of deals meant team revenue figures were no longer shrouded in secrecy, as they had been previously. So while owners celebrated having parlayed successful television broadcasts and distribution contracts, players and their representatives rejoiced in having a definitive figure to target in contract negotiations. That is, owners, who with closed financial books had been able to effectively cry poverty, could no longer do so with any sense of veracity.

In the United States, the NFL has thus far proven to be the most successful sports entity when it comes to consistently maximizing media revenue. But this success has come at the cost of a labor battle that spanned nearly two decades and spilled over into legal disputes to be tried in courts of law. However, one method to achieve labor peace emerged in 1992, when the NFL owners offered to open their financial books and give players a large share of the list of defined revenue streams.

The list was dominated by the league's multibillion-dollar media deal, and the owners offered the players a 63% share (later as much as 65%) in exchange for the players' agreement to cap their salaries at the 63% level. This salary cap, which was similar to the agreement reached a decade earlier within the National Basketball Association (NBA), guaranteed players the lion's share of revenue from the league's media deals and gave the owners cost certainty over their team rosters.

Yet despite the ostensible advantage a salary cap grants to both sides, salary caps have not proven to be as effective in providing cost certainty in all sports. After the NFL and its players association negotiated a collective bargaining agreement in 2006, the prospect of even a slight change in calculation has given rise to speculation that salary caps may not be as effective a tool of cost control as ownership has believed them to be until now. The potential sticking point is that salary caps are tied to overall revenue and contain minimum salary guarantees, and there are a host of other exceptions that have been—and could be even more so going forward—exploited by either ownership or player interests.

But while salary cap controversies and related contract negotiations persist on one hand, on the other hand, for athletes in most major professional sports, both with and without salary caps, compensation is tied to the overall media-earning capacity of their particular teams and leagues. In capped sports, the figures are, by definition, generally fixed; in uncapped sports, the figures are more variable, though ultimately based on media earnings. And in any case, because they have largely signed over their media income to players, owners have turned to new ways of maximizing facility-based revenue streams, a direct result of which is the boom in stadium construction, naming rights, and sponsorship deals during the past 20-plus years.

Owners are, however, not alone in gaining revenue streams through sponsorship and related deals. Since the earliest days of modern professional sports, all manner of corporations and organizations have intended to reach existing and potential customers by aligning themselves with athletes. Professional athletes tend to be able to draw people to an event, which provides the opportunity for organizations to do such things as promote their products and services or motivate and entertain customers and employees. In exchange for making an appearance and performing some relative function, whether in person or through some electronic medium, the athlete typically receives some form of compensation. Player contracts are by and large the primary embodiment of this because athletes possess the status to make their teams and sports more marketable.

Yet over time, especially as the reach of sports and media has grown across the globe, endorsement income has also become an important part of the athlete's overall earnings potential. It is indeed extreme, the case of professional golfer Tiger Woods, who, until derailed by self-destructive behavior that led to personal scandal, was estimated to earn upwards of $100 million per year in endorsement income. But to whatever extent Woods's actions hurt his own earning power and may in the long term impact athletes' endorsements, well-known and recognizable athletes—both current and retired professionals—can earn many millions of dollars annually for endorsing both sport-specific equipment and apparel and consumer products; this is typically in addition to receiving quantities of these items for free.

An athlete's ability to act as a successful endorser is, on the surface, a function of likeability and recognition. But below the surface, the athlete-as-endorser is a tool that is used in an attempt to persuade other individuals to do something in the interest of whatever product, service, or organization is being promoted.

Robert Cialdini of Arizona State University has researched and developed accepted theories about how and why people tend to be influenced by others. Within his framework are the realities that people tend to behave as they see others behaving (so-called social proofs), are likely to outright obey the requests of authority figures, and are easily persuaded by other people they are fond of, especially if they find those individuals somehow attractive. One dominant theme of the literature on the subject is that people are willing to be swayed because it allows them to identify with successful others, which is a means by which to enhance one's self-esteem. That is, when a person joins a renowned group and draws attention to membership in it, that person is likely to feel more satisfied with himself or herself.

Although the theories propounded by Cialdini (1993) are settled in a greater social context, they originate in his and his colleagues' having verified that following a sports team victory, fans are more likely to brandish their team's logo and share the recognition by saying, "We won"; following defeat, the same fans declare, "They lost." These dual concepts—basking in reflected glory and cutting off reflected failure, respectively—help explain the traditional allure of popular and well-behaved athletes as endorsers and the rejection of athletes whose personalities are disagreeable or whose behavior runs afoul of the law or social norms.

Although the basic reasons for using professional athletes to make promotional statements remains the same as anytime before, the advancement of such a strategy and tactic looks very much different today than it did in 1960, when Mark McCormack, founder of International Management Group, began to craft the image of professional golfer Arnold Palmer into one suitable for all manner of endorsements. Palmer was not, of course, the first professional athlete to take on the role of product pitchman. But with McCormack acting as his agent, Palmer is considered to have become the first professional athlete to seriously market, promote, and license himself and his likeness—and to have built a corporate enterprise from it.

The effectiveness of Palmer as a spokesman is a result of advertisers' having chosen to align him with products and services that he either used personally or had confidence in promoting to the public. But the underlying strength of the endorsement strategy rested on and concentrated on something more profound: Palmer as an upstanding human being rather than as a championship golfer. By focusing more on the former than on the latter,

Palmer's value as an endorser was guarded against any likelihood of poor performance on the golf course.

Countless professional athletes across virtually every professional sport have in the interim used the basic tenets of Palmer and McCormack's strategy to convert image into income. And it may well be argued that NASCAR and the teams and drivers within it have individually and collectively built themselves up on much the same premise. What is undeniable, however, is that professional sports have grown to compete variously as part of or side by side with the entertainment industry, and athletes are by now equivalent to entertainers in terms of status, function, and compensation. For high-performing athletes, especially those who have a particular marketability, there are often opportunities to earn more money from endorsing products and services than from playing their sports.

## Global Considerations and Future Directions

In today's global economy, business concerns and information flows are no longer local, regional, or multinational, even if they are organized as such. Business and the flow of information are transnational, which means factors including marketing, pricing, and management do not know national boundaries. But if this analysis is United States–centric, it is only because professional sports in the United States have provided the basic model for the composition, image, and interests of today's professional athletes the world over.

Professional sports teams in the United States, beginning with baseball, have increasingly set about finding players from outside the country. This has been a response, first, to the desire to expand the pool of available talent and, more recently, to increasingly high salaries being paid to players who might not pan out. It is ever more also due to the fantastic attempts by individuals and organizations within the sports business to tap new markets by attracting foreign fans who are conceivably interested in consuming anything that might be related to a favorite professional athlete playing abroad.

With the possible exception of premier football (soccer) players and Formula One race car drivers, both of whom benefit from relative free agency and concomitant auction markets for their services, professional athletes who compete in U.S. leagues are by and large paid a higher salary than their counterparts around the world. This is in great measure due to the victories of their organized labor unions over management. Yet it is fast becoming an old reality.

Although leagues abroad follow the United States in many ways, they tend to have different issues with player movement because they permit the largely unencumbered free flow of players between leagues. In the United States, the existence of one high-professional league per sport means there are a finite number of available roster spots on a team and, by extension, in the league. This phenomenon, combined with the prospects of media exposure, explains, for example, the unprecedented contracts extended to Alex Rodriguez, the first of which doubled the previous record for a sports contract, which was a 6-year, $126 million agreement signed in 1997 between forward Kevin Garnett and the NBA's Minnesota Timberwolves.

The years since the advent of free agency have been a great period of time for the earnings of professional athletes, most of whom have increasingly benefited by attracting owners and sponsors that are willing to compete against one another for the athletes' talents and services. Their status has become more profitable and their access to financially rewarding opportunities has risen steeply. And the farther up the talent and status scale a player has gone, the better the pay. But this much is also true for athletes who compete in individual sports.

Both in and across the team and individual sports labor markets, there is, as in any labor market, inequality of income distribution. The enormous amount of information pertaining to the salaries of professional athletes has permitted reliable Gini coefficients—measurements of inequality in income distributions—to be computed, and the results demonstrate that plenty of inequalities exist. Generally speaking, even in consideration of capped sports, individual versus team sports, and the economic policies specific to each, there is a significant discrepancy between 90% of earners and those in the top 10%, who anyway lag well behind the top 1% (to say little of those in the top 0.1%). Yet given salary increases over time, and despite the few instances of stagnation or decrease, most professional athletes can be assured of receiving better than a living wage.

Consider, for example, that the average salary for an MLB player is at present more than $3 million; when Donald Fehr became executive director of the MLBPA in late 1985, the average player's salary was a little more than $300,000 (approximately $600,000 in 2008 money). The billions of dollars committed to salaries in today's domestic and foreign sports leagues and associations tell nothing of the days when a good many athletes supplemented their incomes by taking jobs during the off-season, which was the impetus for the creation of preseason training camps. Players now typically train year-round and may even receive additional compensation for doing so.

Even so, there is a line of argument that asserts athletes are being exploited because they are being paid wages lower than the MRP, the revenue that is generated. There is, in addition, subtext about the existence of prevailing structures of race and gender discrimination. Although there is not by any means a final word on the matter, it is becoming clear that the argument over whether professional athletes are any longer paid lower wages because they are of ethnic and racial minorities appears to be obsolescent: Alex Rodriguez is Latino; Tiger Woods is of

African American and Asian descent; and NBA players, who are the highest paid of all league athletes in the United States, are a majority black. To be clear, disparities in salary due to racial, ethnic, and even language biases is an existent though decreasing issue in the context of player salaries. However, the same cannot be said for gender. Although female participation in sports has risen dramatically since the latter part of the twentieth century and women's professional leagues have been created in reflection of that change, female athletes are not as well compensated as their male counterparts. One overarching reason for this is because, whatever social stereotypes predominate, professional women's sports have yet to achieve the levels of viewership and, by extension, revenue that men's sports generate.

To be sure, as private investors and public institutions in many parts of the developed world have for the past two decades-plus invested in sports teams and leagues, demand for top players has become increasingly competitive. The U.S.-based leagues have not yet ceased to be the market leaders, but they are beginning to experience realities of the open market that have existed for quite some time almost everywhere else. Most notably, several players in the NBA and the NHL have decidedly gone to play—and be paid handsomely—in elite European leagues. And the implications of acquiring the best talent available has become so fierce that teams in more than a few sports are willing to pay posting and transfer fees across teams and leagues simply for the right to negotiate with a player. Yet no one can say with any certainty whether any of this experience is likely to ring true in years come.

## Conclusion and the Challenges Ahead

Throughout history, professional sports have generally not let outside factors, such as the economy, control business. Executives and personalities have taken charge of external forces, especially during periods of downturn, and acted to endow their assets with various capacities to gain wealth. Having been more proactive than reactive, they have tended to not let economic factors affect business performance any more than those factors set limits on how they conducted their business.

Easy access to credit and other economic factors that existed during the first three quarters of the twentieth century allowed professional sports entities to think and behave as they traditionally had. But significant economic and social transformations in more recent years have led to upheaval of much of what everyone knew to be true. One question that is beginning to be asked is, "Do companies that pay athletes to endorse products and services really benefit from the relationship?"

The existing research is mixed regarding whether athlete endorsements are any longer a defensible corporate decision from a marketing, advertising, or branding standpoint. Now more than ever before, in an era of increased media exposure and player scandal, the company risks gaining negative public attention over the athlete who engages in behavior that is not consistent with the corporate image. But even more profound is that such endorsement deals will be so overused as to be ineffectual, should they continue to proliferate at the rate they have in the past. The problem is, primarily, that companies may well be left with little to distinguish their investments—even if they employ the most renowned athletes and those whose images and values unquestionably connect with the corporate brand. The prospect of this actually happening cannot be dismissed, because collapse eventually emerges in any market in which participants overpay for products or services that do not hold the possibility of producing a fair return on their investments.

Although this and other observations and conclusions outlined in this analysis are based on a wealth of data and literature, there is, altogether, little evidence that today's facts are being interpreted by anything but yesterday's theories. Thus, there is a need to examine and think through the basic assumptions of the sports business and, specific to this discussion, how athletes are compensated for their services.

The first challenge of this effort is to cast a new and functional economic theory for the sports business in general and for player contracts in particular. The formulas for how professional athletes should be compensated for their performance and contributions are often complicated and rather unscientific. But if the proud achievement of billions of dollars being exchanged throughout the sports business is any indication, they have served the interests of both athlete and owner quite well for quite some time. Now given the fits and seizures of the shift to a transnational and knowledge-based economy, the question is, "How much longer can traditional principles and policies be sustained?"

The highest levels of professional sports effectively function as monopolies. As with the majority of monopolistic enterprises, the activity and innovation that begot prosperity was so successful for so long that they find it reasonably difficult to abandon those practices and habits. Thus, if leagues and owners continue to increase the price of admission to games and access to content in order to turn a consistent profit while keeping up with increasing player costs, are they destroying the mutual trust they have with fans? Are the player contracts themselves corrosive because the wages and benefits paid out through them are increasingly so disparate from those of the average fan? Will these circumstances generate waves of contempt that turn athletes who are paid millions of dollars from heroes into villains? If the answers are yes, can players then expect to be paid as much tomorrow as they are today?

Such questions prompt the second challenge, which is to deal with the justification for the terms of player contracts. A long-standing argument has been that professional athletes have a limited window during which to cash

in on their skills. This may well be so; a good case for this could certainly be made for any period prior to free agency. Yet some athletes go so far as to announce that such contracts are necessary if they are to "feed the family." What they mean to say is that the contracts are necessary to feed the family according to a certain standard. This too may be fine, because there is sound argument in paying high-performing individuals according to whatever is deemed an appropriate market value. But in light of today's salaries, benefits, retirement policies, and the access to opportunities afforded to even the least competent professional athletes, the question is whether the rationale matches the reality. When it no longer does (and it may already not), customers—be they fans who attend games or owners who woo players to their teams—reject what is being marketed to them. As a consequence, the entire system lurches toward instability.

A third, but related, challenge is to produce a better and more honest definition of what in the sports business is meant by *short term* and *long term* and what is needed to balance the expectations of both—regardless of whether player contracts are guaranteed. Although much of the subject is beyond this analysis, it nevertheless signals here the need for conscious development of programs that educate and assist professional athletes, whatever their level of financial success and length of professional career, in every aspect of transition to life after their playing days are over. A spectrum of professional athletes, from the extremely well known to those who appeared in the ranks momentarily, have found themselves impoverished by wrestling with the frustrations of trying to find post-career outlets that in even a small way replicate the sense of competition, camaraderie, and notoriety associated with being a professional athlete.

These are not, of course, the only challenges facing the incomes and earnings of professional athletes at the outset of the twenty-first century. But in considering the rapid and impressive changes to athletes' salary and endorsement prospects, these challenges are certainly high among the list of priorities going forward.

## References and Further Readings


Berri, D., Schmidt, M., & Brook, S. (2006). *The wages of wins: Taking measure of the many myths in modern sport.* Stanford, CA: Stanford University Press.

Cialdini, R. (1993). *Influence: The psychology of persuasion.* New York: Quill William Morrow.

Coakley, J., & Dunning, E. (2002). *Handbook of sports studies.* Thousand Oaks, CA: Sage.

Drucker, P. (1993). *Post-capitalist society.* New York: HarperCollins.

Helyar, J. (1994). *Lords of the realm: The real history of baseball.* New York: Villard Books.

Lewis, M. (2003). *Moneyball: The art of winning an unfair game.* New York: W. W. Norton.

MacCambridge, M. (2004). *America's game: The epic story of how pro football captured a nation.* New York: Random House.

Miller, M. (1991). *A whole different ball game: The sport and business of baseball.* New York: Birch Lane.

Palmer, A., & Dodson, J. (1991). *A golfer's life.* New York: Random House.

Rosner, S., & Shropshire, K. L. (2004). *The business of sports.* Sudbury, MA: Jones & Bartlett.

Rottenberg, S. (1956). The baseball players' labor market. *Journal of Political Economy, 64*(3), 242–260.

Thaler, R. H. (1992). *The winner's curse.* Princeton, NJ: Princeton University Press.

Zimbalist, A. (1992). *Baseball and billions.* New York: Basic Books.

# 54

# Economics of Gender

Lena Nekby
*Stockholm University*

Peter Skogman Thoursie
*Institute for Labour Market Policy Evaluation (IFAU), Uppsala*

This chapter on the economics of gender discusses the role of gender in the labor market from an economic perspective. Focus on the labor market is motivated by the fact that labor earnings are arguably the most important component of an individual's income and a major determinant of living standards. Earnings are also correlated with employment opportunities, occupation, promotion, and job mobility, all of which, for reasons that are discussed in greater detail below, are influenced by gender. Although there are other arenas where gender is of economic importance (e.g., gender differences in the division of child care and domestic labor, access to day care, and health), these topics are not covered in this chapter. Interested readers should instead turn to the chapters on the economics of the family and feminist economics. This chapter focuses on the theoretical foundation for gender differences in the labor market as well as the empirical evidence on gender discrimination and gender differences in preferences.

The chapter begins with an overview of country differences in gender wage and employment gaps. This section discusses how gender differences in employment rates and differences in wage dispersion may relate to the gender wage gap. This is followed by a theoretical overview of gender differences in the labor market covering both supply-side differences in human capital acquisition—the role of gender-specific preferences—and demand-side discrimination. The next section discusses the empirical problem of identifying discrimination in the labor market. The chapter rounds off with two sections covering, in turn, the empirical evidence on discrimination and gender differences in preferences.

## Gender Differences in the Labor Market: Overview

Studying differences in labor earnings is of fundamental importance for anyone interested in understanding poverty, social stratification, and the economic incentives facing workers. Labor earnings are the most important component of an employed individual's income and a major determinant of living standards. This chapter therefore begins with a comparison of gender labor market gaps in a selected number of industrialized countries. First, an overall picture of gender wage gaps and differences between countries is provided, followed by a description of the trends in these gaps during the last 20 years. Country differences in gender wage gaps are related to differences in female employment rates as well as varying wage distributions across countries. Possible explanations for the wage gaps observed, such as productivity differences, occupational gender segregation, preferences, and discrimination are discussed in subsequent sections.

Table 54.1 reports the gender wage gap, the female employment rate, and the fraction of female employees who work part-time for a selected number of countries during the period from 1989 through 2006. The gender wage

gap is defined as 1 minus the ratio of the annual averages of female and male mean hourly wage rates, which can then be interpreted as how much less, as a percentage, women earn per hour relative to men. For example, a gender wage gap of 23% in the United States in 2006 means that the female wage rate is 23% smaller than the male wage rate. In other words, for every $100 that men earn per hour, women earn 23% less—that is, $77. As shown in Table 54.1, gender differences in pay prevail in all countries, even though the size of the gaps varies considerably across countries. In 2006, the female hourly wage rate in France was 11% smaller than the male hourly wage rate. Thus, wages of employed women in France are closer to men's wages than in the case of the United States. When one looks at the period from 1994 through 1998, where information on the gender wage gap for the full set of countries is available, one sees that the gender wage gap is 36% in Japan, 10% in France, and 24% in the United States. Table 54.1 also shows that there seems to be a tendency toward decreasing gender wage gaps over time in most countries since the late 1980s.

An interesting question to ask is why gender differences in labor market outcomes vary across countries. For example, why do women earn less relative to men in a country such as the United States compared with France or Sweden? One reason could be that women have acquired different skill levels across countries and that women in France and Sweden are more qualified than women in United States. Because skills are rewarded through higher pay in the labor market, this is one potential reason for observed country differences in pay. According to Francine Blau and Lawrence Kahn (2000), however, there seems to be little reason to believe that women in the United States are less qualified relative to men than women in other countries. An alternative explanation is that there are country differences in economic returns to skills and therefore differences in economic incentives. Countries with high rewards to skills have wage structures that encourage skill acquisition among workers. This suggests that the wage structure in a country plays an important role in determining the gender wage gap, given that there are gender differences in skills and qualifications. Consider, for example, two countries where women have lower levels of labor market experience than men, but the gender difference in experience is the same in the two countries. If the return to experience is higher in one country, this country will have a larger gender wage gap, all else equal.

Moreover, centralized wage-setting institutions tend to reduce wage dispersion across firms and industries and raise the relative pay of low-wage workers (regardless of gender), which in turn may reduce the gender wage gap. Because most European countries have more centralized wage setting compared with the United States, the degree of centralization of wage setting may be an important explanatory factor behind country differences in the gender wage gap. Empirical evidence suggests that the overall wage structure is of major importance in explaining gender wage gaps, where the higher level of wage inequality in the United States compared with other countries tends to increase U.S. gender differentials relative to those in other countries (see, e.g., Blau & Kahn, 1996a, 1996b).

These explanations of the gender wage gap are based on the assumption that women on average are less qualified than men. It is therefore interesting to see how the gender wage gap has evolved over time as female labor force participation has increased in most countries since the 1970s. Human capital skills, such as level of education and labor market experience, have also increased for females relative to males during this period. Because a stronger attachment to the labor force would increase other labor market skills as well, this suggests that gender wage gaps should shrink over time. According to Table 54.1, the female employment rate has increased slightly in most countries, together with a decrease in the gender wage gap. Thus, it appears that women to some extent have caught up with men by accruing more skills. It should be stressed, however, that even if women to a large extent have caught up with men in terms of level of education, they systematically chose different types of educations. For example, men to a larger extent than women chose technical educations. If technical educations and the occupations associated with these educations yield higher returns in the labor market than other types of educations, this would improve men's labor earnings relative to those of women.

Furthermore, because women have lower employment rates than men, women might be a more selected group in the labor market than men with, on average, higher tastes and skills for work compared to the population of women. Such selection into employment makes it difficult to study trends in gender wage gaps because the group of female workers from one time period to another might not be comparable when female employment rates change over time. This also makes across-country comparisons of gender wage gaps difficult because female employment rates are very different across countries, implying that the selection of females who participate in the labor market in one country is potentially very different from the corresponding selection in another country.

Difficulties in comparing trends in gender wage gaps across countries are supported by the striking international variation in female employment rates. It is apparent in Table 54.1, for example, that France had the lowest gender wage gap in 2006 (11%) but also one of the lowest female employment rates. Only 59% of women were employed in France at that time, compared with 66% of women in the United States. One hypothesis, therefore, is that selection into employment is not random and might affect the size of gender wage gaps. In particular, if employed women have relatively high-wage characteristics, low female employment rates are consistent with lower gender wage gaps because women with low-wage characteristics are not included in the observed wage distribution. This may explain the negative correlation between gender wage and employment gaps that are observed in Table 54.1.

The pattern of countries with high gender wage gaps tending to have high female employment rates might also

**Table 54.1** Gender Wage Gaps, Female Employment Rates, and Female Part-Time Work Rates Among the Employed (Percentages)

| | *1989–1990* | *1994–1998* | *2006* |
|---|---|---|---|
| **Austria** | | | |
| Gender wage gap | 33 | 31 | — |
| Female employment rate | — | 59 | 64 |
| Female part-time work rate | — | 30 | 40 |
| **France** | | | |
| Gender wage gap | 15 | 10 | 11 |
| Female employment rate | 52 | 53 | 59 |
| Female part-time work rate | 25 | 32 | 30 |
| **Italy** | | | |
| Gender wage gap | 20 | 17 | — |
| Female employment rate | — | 37 | 46 |
| Female part-time work rate | — | 14 | 26 |
| **Japan** | | | |
| Gender wage gap | 41 | 36 | — |
| Female employment rate | 57 | 57 | 59 |
| Female part-time work rate | — | — | — |
| **Sweden** | | | |
| Gender wage gap | 21 | 16 | 16 |
| Female employment rate | 73 | 68 | 71 |
| Female part-time work rate | — | 34 | 40 |
| **United Kingdom** | | | |
| Gender wage gap | 32 | 25 | 21 |
| Female employment rate | 61 | 64 | 66 |
| Female part-time work rate | 44 | 44 | 42 |
| **Unites States** | | | |
| Gender wage gap | 29 | 24 | 23 |
| Female employment rate | 64 | 67 | 66 |
| Female part-time work rate | — | — | — |

NOTES: Female part-time work rates are the fraction of employed women who work part-time. Female employment rates and female part-time work rates are from Eurostat (2009). Gender wage gaps are based on hourly wage rates. Gender wage gaps for the periods from 1989 through 1990 and 1994 through 1998 are taken from Blau and Kahn (2000), and the gender wage gaps in 2006 are from Eurostat. The gender wage gap for the United States in 2006 is based on annual earnings for full-time workers (Institute for Women's Policy Research, 2009).

be reinforced by differences in the extent to which women work part-time. In France, the fraction of employed females working part-time in 2006 was 30%. The corresponding figure for Sweden, which had a larger gender wage gap as well as a higher female employment rate in 2006, was 40%. Working part-time implies that a smaller amount of work experience is accumulated over time in comparison to full-time workers. It might also be the case that working part-time is associated with lower chances for promotion and wage raises. Taken together, there are, therefore, a number of explanations as to why a higher female employment rate, together with high female part-time rates, might be associated with a higher gender wage gap within a country.

Different patterns of employment selection across countries may in turn stem from a number of factors. First, there may be country differences in the gender role of household work, social norms, or both, affecting labor force participation. Second, labor demand mechanisms, including social attitudes toward female employment and those attitudes' potential effects on employer choices, may be at work, affecting both the employment rate as well as the level of wage offers by gender. Claudia Olivetti and Barbara Petrongolo (2008) suggest that the international variation in gender employment gaps can indeed shed light on across-country differences in gender wage gaps. This study suggests that sample selection into employment explains nearly one half of the observed negative correlation between gender wage and employment gaps. They also show that while the raw wage gap is much higher in Anglo-Saxon countries than in southern Europe, the reason is probably not to be found in more equal pay treatment for women in the latter group of countries but rather in different selection processes into employment. Female participation rates in southern European countries are low and concentrated among high-wage women. Correcting for lower participation rates in southern European countries widens the wage gap to levels similar to those of other European countries and the United States.

## Theoretical Explanations for Gender Differences in the Labor Market

Gender differences in labor market outcomes stem from supply-side differences in productivity, labor supply, or preferences or to demand-side differences in opportunity—that is, gender discrimination in employment, wage-setting, or promotion of equally qualified individuals (see Altonji & Blank, 2003, for an overview). Supply-side differences in productivity and labor supply between men and women are often analyzed within the human capital framework (Mincer & Polachek, 1974). The human capital model postulates that individuals invest in education and training and are rewarded for these investments in the labor market, either via enhanced employability, higher wages, or both (Mincer, 1958; Mincer & Polachek, 1974). Within this framework, gender differences in labor market

outcomes are attributable to gender differences in human capital investment.

A number of explanations have been forwarded as to why women historically have invested less in education, skills, and other qualifications valued in the labor market. A partial explanation can be found in traditional gender norms concerning child care and housework within families. If women expect to spend more time out of the labor market, they consequently have less time in the labor market to reap the benefits of their human capital investments. As such, women will invest less in market-oriented human capital and, in addition, will direct their investments toward educations and skills with lower depreciation rates due to time out of the labor market (Polachek, 1975). Notice that gender differences in the types of academic training being invested in, due to differential depreciation, also help to explain patterns of occupational segregation by gender. Shorter duration in the labor market as well as more frequent interruptions for child rearing also influence employers' willingness to finance on-the-job training for female employees as well as promotion possibilities, reinforcing, over time, gender differences in productivity. Note that there is a literature that questions the degree to which human capital models can explain occupational segregation by gender (Beller, 1982; England, 1982).

Not only are there social norms concerning women's division of labor between home and market production but there are also norms concerning what constitutes typically female or male pursuits in the labor market (Akerlof & Kranton, 2000). In other words, preferences for certain types of occupations or educations may be due to social norms or preconditioning regarding what is considered appropriate for women and the costs of deviating from these norms (Akerlof & Kranton, 2000; Gundersson, 1989; Polachek, 1984). Differences in type of academic training can also be the result of historical differences in remuneration and access to jobs by gender. If women historically have had lower access to (due to norms or discrimination) or lower payoffs from certain occupations, then investments in education and training for these occupations will also be lower. This implies that historical discrimination in the labor market can lead to subsequent differences in human capital investment and that gender differences in the labor market can become a self-fulfilling prophecy (Darity & Mason, 1998). Claudia Goldin (2006) argues that lower remuneration in typically female occupations has an historical basis in the segregation of jobs as women increasingly entered the labor market in the early 1900s. As white-collar positions opened up for women, policies were instituted at the firm level, creating sex-segregated positions. Jobs were increasingly classified as either female or male, where the majority of female jobs were dead-end, providing little room for advancement to higher positions or earnings growth.

Over time, shifts in the norms concerning female participation in the labor market, greater and more continuous time in the labor market, as well as more equitable distribution of housework and child care will increase incentives for women to invest in human capital accumulation and for firms to invest in female employees. Indeed, the salience of gender differences in human capital investment as an explanation for gender differences in labor market gaps has decreased over time as women increasingly have closed the gender gap in education, at least with regard to level of education. However, hard-to-break differences in remuneration attributable to occupational segregation remain due to either cultural devaluation of female jobs or dual labor market and crowding theories, where exclusion from male jobs leads to crowding in female jobs and consequently lower wages as well as lower returns to education (Bergmann, 1974; Doeringer & Piore, 1971). Occupational segregation by gender is therefore both a cause and a consequence of gender differences in pay. Occupational segregation can lead to gender differences in human capital investment and productivity, but it is also a partial explanation for gender wage and income differentials in a society characterized by occupational gender segregation and lower remuneration for female jobs.

On the demand side, two forms of discrimination are commonly discussed within the economics framework: taste-based discrimination and statistical discrimination. Taste-based discrimination arises because of a disutility among employers (customers or coworkers) for interacting with female workers or because of a preference for male workers (Becker, 1971). If tastes for discrimination are large and the demand for preferred male workers is lower than the supply, a wage differential arises between male and female workers, the implication of which is a competitive advantage for firms that hire equally productive women. This suggests that wage gaps will disappear in the long run as nondiscriminating employers enter the market. Frictions to free entry, imperfect information, collective bargaining, search costs, and other infringements to perfect market competition may, however, lead to sustainable wage gaps due to taste-based discrimination over time. There are a number of empirical studies examining the degree to which competition decreases gender discrimination; see, for example, Orley Ashenfelter and Timothy Hannan (1986); Sandra Black and Elizabeth Brainard (2004); Black and Philip Strahan (2001); Judith Hellerstein, David Neumark, and Kenneth Troske (2002); Xin Meng (2004); and the references therein.

Statistical discrimination arises because of the inability to acquire, or costs of acquiring, perfect information about job candidates. Instead, employers use readily available group statistics to assess candidates (Arrow, 1973; Phelps, 1972). This implies that if women on average have lower relevant labor market experience for a given position, are expected to leave the labor market for child rearing, or both, to a larger extent than male candidates, employers will be less inclined to interview and hire female candidates. Notice that statistical discrimination is individual

discrimination based on actual group statistics. The individual in question may deviate from mean group characteristics, but employers are unable to assess this information, or the costs of doing so are high. Statistical discrimination is also based on the assumption that unbiased group statistics are readily available. This may not be the case, and agents may act on the presumption that their beliefs are statistically correct but actually base decisions on biased information because of erroneous perceptions of group productivity. In the sociological literature, this is termed *error discrimination* (England, 1992).

## The Empirical Problem of Identifying Gender Discrimination in the Labor Market

The most common method of estimating gender differences in labor market outcomes is through wage, earnings, or employment regressions using register or survey data. Typically, variations of the following basic model are estimated:

$$y_i = \alpha + \beta_1 female_i + X_i\beta_2 + \varepsilon_i$$

where $y_i$ is a labor market outcome (employment, wages, income) of individual *i, female* is a binary (dummy) variable equal to 1 if female and 0 otherwise, and $X$ is a vector of other control variables, typically both demographic and human capital indicators that may influence the outcome variable and systematically differ between men and women. Finally, $\in$ is the random error component measuring the impact of unobserved characteristics in the equation that also influence the outcome variable. In estimation of this type of equation, the coefficient for *female* indicates to what degree labor market outcomes differ between men and women with similar characteristics (the observable characteristics included in the vector $X$). It does not, however, tell us whether these gender differences are due to discrimination—that is, to unequal treatment of equally qualified individuals. The reason is that it is impossible for researchers to control for all possible supply-side differences in productivity by gender. For a causal interpretation of the effect of gender on a given outcome, the correlation between the *female* dummy variable and the random error component must be equal to 0. This is called the *zero conditional mean assumption,* stating that there should be no systematic differences in unobserved factors between men and women that influence the outcome of interest. If there is such a factor—for example, if women have systematically lower on-the-job training than men and this is unobserved in an income equation—then the measure for gender differences in income will be biased if differences in labor market training are not accounted for in estimation. This implies that observed gender differences in labor market outcomes may be due to supply-side differences in productivity by gender, uncontrolled for in estimation (unobserved characteristics) or discrimination.

Note that there is a large literature concerning the decomposition of gender wage gaps into an explained and unexplained component, the so-called Oaxaca-Blinder method, where the unexplained component measures wage differentials that remain after controlling for all observable characteristics in estimation (Blinder, 1973; Oaxaca & Ransom, 1999). (For extensions of this method, see also Juhn, Murphy, & Pierce, 1991.)

Recently, attention has turned to estimating gender differences in the labor market using experimental methods. The reason is that a well-executed controlled experiment can unequivocally identify a causal effect of gender on an outcome of interest. Experiments are based on random assignment of a given population into a treatment group and a control group. Random assignment guarantees that both observable and unobservable individual characteristics are, on average, similar in treatment and control groups. This implies that any differences in outcomes due to treatment are attributable to gender alone. Studies based on experimental methods are often characterized by high internal validity but low external validity. High internal validity implies that experiments are credibly able to identify a causal effect of gender on outcomes of interest. Low external validity implies that results often cannot be generalized to the population at large. This is due to the fact that experiments are often performed on limited subsamples of the population only. Experimental methods have been used to study the presence of gender differences in preferences as well as employer discrimination. The next two sections provide an overview of the recent empirical evidence within these two strands of research on gender and the labor market.

## Evidence of Gender Discrimination

Two commonly used methods to study gender discrimination in the labor market are so-called audit and correspondence testing studies (Darity & Mason, 1998; Pager, 2007; Riach & Rich, 2002). Audit studies use actual test persons to apply for jobs and participate in job interviews, while correspondence testing studies send written applications (CVs) to actual job openings and measure differences in callbacks to interviews. Audit studies have the advantage of testing discrimination in the entire application process, from initial contact with employers to actual job offers. On the other hand, audit studies are unable to credibly control that test persons are similar in all characteristics of relevance to employers (Heckman, 1998; Heckman & Siegelman, 1993). Although test persons are trained to behave and dress in a similar manner, their use introduces uncertainty into the experiment. Female test persons may, for example, internalize expected discrimination and subconsciously behave differently than male test persons. These issues are avoided in correspondence testing studies, where interpersonal contact between test subjects and employers

is avoided via the use of written applications only. Many studies also suggest that the majority of employer discrimination occurs at this stage—that is, in callbacks to interviews from formal applications (Riach & Rich, 2002). As such, the correspondence testing methodology is equivalent to a randomized controlled experiment where gender is signaled by, for example, the names assigned to CVs.

The first field experiment to study gender discrimination in hiring was carried out by Peter Riach and Judith Rich (1987) in Victoria, Australia, using the correspondence testing methodology. This study finds that women encountered discrimination 40% more often than men. In a more recent work, Neumark (1996) studies the prevalence of gender discrimination in restaurant hiring in Philadelphia (United States) using both audit and correspondence testing methods. Female and male pairs were matched and trained to apply for waitress and waiter jobs in low-, medium-, and high-price restaurants after providing written CVs to restaurant managers. This experiment design therefore metes out possible differences in test person behavior by testing employer responses to both written applications and job interviews. Results showed that high-priced restaurants were significantly more likely to interview men than women (gender difference in callbacks to interviews) and significantly more likely to offer men jobs after interviews. Gender differences in hiring were not significant in low- and medium-priced restaurants. As wages and tip earnings are significantly higher in high-priced restaurants relative to low- and medium-priced restaurants, hiring discrimination in high-priced restaurants provides a partial explanation for within-occupation gender wage differentials in the restaurant sector.

Other correspondence tests of gender discrimination in hiring attempt to explicitly link hiring discrimination to statistical discrimination, either via experiments set up to test the effect of employer expectations concerning forthcoming childbirth among young female applicants (Duguet & Petit, 2006) or by signaling unobserved stereotypical personality traits associated with gender (Weichselbaumer, 2004). Emmanuel Duguet and Pascale Petit (2006) find that for qualified job positions, younger childless women have lower callbacks to interviews than men. This difference vanishes for older women, regardless of whether they have children. Doris Weichselbaumer (2004) tests whether women experience less hiring discrimination in traditionally male jobs if they signal stereotypical male personality traits (ambition, assertiveness, competitiveness, dominant individualism), which are normally not observed on written CVs but are nonetheless assumed to be important for the hiring decision. Three applications were sent to each job opening: two female and one male, where the male applicant and one of the female applicants signaled typically male personality traits while the other female applicant signaled traditionally female personality traits. Results showed that women were treated equally, regardless of variation in personality traits. Unfavorable treatment for women applying to masculine occupations was not mitigated for women signaling masculine traits, and preferential treatment for women in feminine occupations was not threatened by masculine traits among female job candidates.

More recent field experiments attempt to link occupational segregation and hiring discrimination. Riach and Rich (2006) use the correspondence testing methodology to test for sexual discrimination in hiring in England, explicitly testing occupations that can be categorized as typically female, male, or mixed. Men were found to be discriminated against in the female occupation (secretary), and women in the male occupation (engineer). However, and somewhat unprecedented in the research literature to date, significant discrimination against men was found within mixed occupations (trainee chartered accountant and computer analyst programmer). Mangus Carlsson and Dan-Olof Rooth (2008) test 13 occupations in Sweden with varying gender composition and find no hiring discrimination against women in male-dominated occupations and a slight preference for women in female-dominated occupations as well as mixed occupations. This study therefore suggests that present hiring discrimination is not likely to explain substantial occupational segregation by gender.

Another innovative field study uses the introduction of blind auditions to analyze potential hiring discrimination against female musicians within symphony orchestras (Goldin & Rouse, 2000). In the 1970s and 1980s, many major U.S. orchestras changed their audition policies and adopted so-called blind auditions—that is, auditions behind a screen preventing the identification of applicants' gender. Interestingly, many auditions took the added precaution of muffling the sound of footsteps, which may otherwise have betrayed the gender of candidates, either by rolling out a carpet to the stage or by asking candidates to remove their shoes. Results from this study indicate that the adoption of a screen increased by 50% the probability that a female musician advanced beyond preliminary rounds and by several-fold the likelihood that a female musician would win in the final round.

A recent study on discrimination in hiring attributable to the gender composition of evaluation committees ties together the two strands of research discussed in this chapter—namely, gender differences in preferences and discrimination. Manuel Bagues and Berta Esteve-Volart (2008) use unique evidence provided by the public examination of candidates applying for positions within the Spanish Judiciary. Candidates are randomly assigned to evaluation committees with varying gender composition. Results from this study indicate that female candidates are less likely to be hired and male candidates more likely to be hired when randomly allocated to a committee with a larger share of female evaluators. A careful analysis of the data suggests that the quality of male candidates is overestimated in committees with female majorities. If women systematically overestimate the quality of male

candidates, this could partially explain female reluctance to compete, as well as relatively poor performance in competitive environments, as noted in several studies that are discussed in the next section on gender differences in preferences.

The studies discussed in this section have spawned a large research interest in testing the presence of gender discrimination in different countries, labor markets, and populations. The consensus in this literature is that gender continues to be a salient factor in hiring within many labor markets and that more research is necessary to understand the extent to which discrimination explains labor market gaps today as well as the extent to which discrimination influences the preferences and human capital acquisition of coming generations.

## Evidence on Gender Differences in Preferences

There is a vast literature aimed at explaining gender differences in preferences. Gender differences in preferences are interesting in and of themselves but also have potential importance in explaining gender labor market gaps. Men and women might have different preferences with respect to risk taking and competition, as well as different reactions to competition. For a survey of the research literature on gender differences in preferences, see Rachel Croson and Uri Gneezy (2009). Different degrees of risk taking among individuals can translate into different choices concerning jobs and occupations, according to the individuals' risk exposure to unemployment, work injuries, and so forth. If workers are compensated for such risks through higher earnings, one explanation for observed gender wage gaps would be gender differences in risk taking. If women are less likely to compete, it not only reduces the number of women who enter competitive environments—for example, in terms of competition for jobs or wage bargaining—but it also decreases the chances for women of succeeding in these competitions.

The general finding from studies on gender differences in risk-taking behavior, most of which are based on laboratory experiments, is that men are more prone to risk taking than women. There are several explanations as to why men are more risk-prone than women. A number of studies, for example, find men to be more overconfident than women. Muriel Niederle and Lise Vesterlund (2007) test this using a controlled laboratory experiment where pairs of two women and two men perform a task, namely adding up sets of five 2-digit numbers for 5 minutes. Participants were first asked to perform the task with piece-rate compensation and thereafter in a tournament setting. After having experienced both compensation schemes, participants were asked to choose one of the two payment schemes in the final round (either piece rate or tournament). Although there were no gender differences in performance in either compensation scheme, Niederle and Vesterlund find that 73% of men prefer the tournament scheme compared with only 35% of women. To elicit participants' beliefs on their relative tournament performance, at the end of the experiment, participants were asked to guess how their performances ranked relative to those of other participants. While 75% of men thought they were best in their group, only 43% of women held this belief. Participants who were more confident about their relative tournament performance were also more likely to enter a tournament.

Although numerous studies such as this have found men to be more overconfident than women on average, this difference may not hold for all tasks or for selected participants. One example of this is provided by Lena Nekby, Peter Skogman Thoursie, and Lars Vahtrik (2007), who use a large running race to study the behavior of women who choose to compete in a male-dominated setting and the consequences of this behavior on performance. Participants were given the opportunity to self-select into start groups based on individual assessments of running times. Overconfident behavior was measured as self-selection into start groups with lower time intervals than final results in the same race (or in the previous year's race) would motivate. Only runners who participated in the same race on the same course during the previous year were sampled so that results would not be contaminated by potentially lower knowledge of individual capabilities among women. Results show that there are environments (male dominated) in which the selection of women who participate are more likely to be confident and competitive and that, within this group, performance improves equally for both genders.

This result is important because gender differences in labor outcomes may be underestimated in selective environments, such as among executives. Earlier studies on the gender wage gap, for example, have found a glass ceiling for women in the upper part of the income-wage distribution (see, e.g., Albrecht, Björklund, & Vroman, 2003). One interpretation of a glass ceiling is that women have greater difficulties than men in obtaining higher positions for observationally equivalent qualifications due to unobservable differences in competitiveness. If there are women in male settings who are as competitive as men, women who compete for higher positions may be evaluated on average female behavior and statistically discriminated against and prevented from reaching higher positions. Another example of this is found among professional investors. Although men are more risk taking than women on average, there are no differences in risk propensities among professional investors, as indicated by their investment behaviors (see, e.g., Atkinson, Baird, & Frye, 2003). Thus, while fewer women are selected into positions such as investors and executives, those who do choose to enter these professions have similar risk preferences to the men in these positions.

In terms of gender differences in competitive behavior, men seem to choose competitive environments to a larger

extent than women. In contrast to Niederle and Vesterlund (2007), there is also some evidence that men perform better than women in competitive environments. For example, in a study by Gneezy, Niederle, and Aldo Rustichini (2003), men and women were asked to solve mazes on a computer for 15 minutes. Participants were paid either according to a piece rate (a dollar amount per maze solved) or according to a winner-take-all tournament. Under the piece-rate scheme, no significant differences in performance between men and women were found. When participants were paid on a competitive basis, however, the performance of men significantly increased, relative to the performance of women. However, similar to this result concerning selective participation, there is evidence that women who choose to compete perform as well as men in these settings (see, e.g., Datta Gupta, Poulsen, & Villeval, 2005).

In a field study, Gneezy and Rustichini (2004) study children's running performance in a regular school physical education class. Children were asked to run twice over a short track while the teacher measured their speed. First, the children ran alone. Thereafter, the children were paired according to running times. This led to pairs of runners with different gender compositions. When the children ran alone, no gender differences in performance were noted. However, when the children ran in pairs—that is, in competition, running times for boys improved by 0.163 seconds on average, while the performance for girls decreased by 0.015 seconds relative to the when they ran alone. Boys were apparently spurred by competition to a larger extent than girls. One objection to these types of studies is the degree to which the produced results can be generalized to actual labor market settings. A similar concern is the extent to which results are task or location specific.

One area where competition has a potentially strong impact on labor market outcomes is in situations concerning bargaining. Deborah Small, Michele Gelfand, Linda Babcock, and Hilary Gettmen (2007) study bargaining using a laboratory setting where participants were told in advance that they would be paid between $3 and $10 for participation at the end of the experiment. After the participants finished their assigned tasks, the experimenter thanked them for their participation and asked, "Here is three dollars. Is three dollars OK?" Only 2.5% of female participants but 23% of male participants requested more money. In another study, Jenny Säve-Söderberg (2007) uses data from two surveys of recent law graduates on their transitions from school to work. Results in this study indicate that in wage negotiations, women consistently submitted lower wage bids than men. It was also found that women received lower wage offers, even when women and men submitted the same wage bid. Negotiating for wages was, however, found to lead to higher wage offers than not submitting any wage bid at all.

The discussion thus far has pointed out that there seems to be evidence of gender differences in preferences, but no discussion on the possible reasons why there might be differences in preferences. Explanations for observed gender differences in preferences include possible genetic differences (see Colarelli, Spranger, & Hechanova, 2006). In contrast, Gneezy, Kenneth Leonard, and John List (2009) use an experimental approach to explore whether there are gender differences in selecting into competitive environments across cultures by examining a patriarchal society (the Maasai in Tanzania) and a matrilineal society (the Khasi in India). Similar to the evidence presented for Western societies, Maasai men opt to compete at twice the rate as Maasai women. The opposite result is found among the Khasi, where women choose the competitive environment considerably more often than men. These results suggest that existing societal structures are linked to observed gender differences in competitiveness. Croson and Gneezy (2009) summarize this literature and conclude that there is support for both genetic and cultural explanations for gender differences in competition. The interesting question for further research is what weight to attach to each of these factors in explaining gender labor market gaps.

Taken together, results from studies on gender differences in preferences suggest that men on average are more inclined to take risks than women and that men also prefer competitive environments to a larger extent. These findings potentially explain a portion of observed gender differences in the labor market because risk taking and competitive behavior are likely to be associated with pecuniary awards. To what degree observed gender differences in preferences are genetic or socially determined is an open question for further research.

## Conclusion

This chapter has discussed the role of gender in the labor market from an economic perspective. Theoretically, there are potential supply-side differences by gender in skill acquisition and preferences that can explain why women earn less on average than men. On the other hand, there are also potential gender differences in opportunity—that is, discrimination prior to entering the labor market, affecting skill acquisition and working norms, as well as direct discrimination in the labor market, affecting employment and earning. Employers may, for example, expect women, to a larger extent than men, to have longer and more frequent absences from the labor market. Women as a group may then be perceived as a greater risk by employers then men, leading to lower wages and promotion possibilities and also to lower incentives to invest in human capital for future generations.

Evidence from studies on gender discrimination suggests that unequal treatment of equally qualified men and women continues to be important in the labor market. Experimental studies on discrimination tend, however, to be carried out on limited subsamples of the labor market,

implying that it is difficult to make any general conclusions about the extent to which discrimination can explain gender gaps in the labor market.

Studies on gender differences in preferences, in turn, suggest that men on average are more inclined to take risks than women and that men also prefer competitive environments to a larger degree than women. These findings potentially explain a proportion of gender differences in the labor market because risk taking and competitive behavior are likely to be positively awarded by employers.

The consensus in the research literature on gender and economics is that both preferences and discrimination matter for labor market outcomes, outcomes that are of crucial importance for lifetime income, living standards, and intergenerational transmissions of income and opportunity, implying that attention must be paid to minimizing the impact these two overriding factors have in creating gender labor market gaps. There are also social changes in force that with time are likely to alter the underlying mechanisms behind these gender differences; among them, changing norms concerning the within-family allocation of labor between household and market, access to day care, and legislation promoting equal opportunity. The question remains to what degree social changes alone will eliminate differential opportunities by gender and to what degree more active measures are necessary to enforce equal opportunity.

## References and Further Readings


Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics, 115*(3), 715–753.

Albrect, J., Björklund, A., & Vroman, S. (2003). Is there a glass ceiling in Sweden? *Journal of Labor Economics, 21*(1), 145–177.

Altonji, J. G., & Blank, R. (2003). Race and gender in the labor market. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3C, pp. 3143–3260). New York: Elsevier North-Holland.

Arrow, K. (1973). The theory of discrimination. In O. A. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton, NJ: Princeton University Press.

Ashenfelter, O. A., & Hannan, T. (1986). Sex discrimination and product market competition: The case of the banking industry. *Quarterly Journal of Economics, 101*(1), 149–174.

Atkinson, S. M., Baird, S. B., & Frye, M. B. (2003). Do female fund managers manage differently? *Journal of Financial Research, 26,* 1–18.

Bagues, M. F., & Esteve-Volart, B. (2008). *Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment* (FEDEA Working Paper No. 2007-15). Madrid, Spain: FEDEA.

Becker, G. (1971). *The economics of discrimination* (2nd ed.). Chicago: University of Chicago Press.

Becker, G. (1993). *Human capital: A theoretical and empirical analysis, with special reference to education* (3rd ed.). Chicago: University of Chicago Press.

Beller, A. H. (1982). Occupational segregation by sex: Determinants and changes. *Journal of Human Resources, 17*(3), 371–392.

Bergmann, B. (1974). Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal, 1*(1–2), 103–110.

Black, S. E., & Brainard, E. (2004). Importing equality? The impact of globalization on gender discrimination. *Industrial and Labor Relations Review, 57*(4), 540–559.

Black, S. E., & Strahan, P. E. (2001). The division of spoils: Rent sharing and discrimination in a regulated industry. *American Economic Review, 91*(4), 814–831.

Blau, F. D., Brinton, M. C., & Grusky, D. B. (Eds.). (2006). *The declining significance of gender.* New York: Russell Sage.

Blau, F. D., & Kahn, L. M. (1996a). International differences in male wage inequality: Institutions versus market forces. *Journal of Political Economy, 104*(4), 791–837.

Blau, F. D., & Kahn, L. M. (1996b). Wage structure and gender earnings differentials: An international comparison. *Economica, 63,* S29–S62.

Blau, F. D., & Kahn, L. M. (2000). Gender differences in pay. *Journal of Economic Perspectives, 14*(4), 75–99.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources, 8,* 436–455.

Carlsson, M., & Rooth, D.-O. (2008). *An experimental study of sex segregation in the Swedish labour market: Is discrimination the explanation* (IZA Discussion Paper No. 3811). Bonn, Germany: Institute for the Study of Labor.

Colarelli, S. M., Spranger, J. L., & Hechanova, M. R. (2006). Women, power, and sex composition in small groups: An evolutionary perspective. *Journal of Organizational Behavior, 27,* 163–184.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature, 47*(2), 448–474.

Darity, W. A., & Mason, P. L. (1998). Evidence on discrimination in employment: Codes of color, codes of gender. *Journal of Economic Perspectives, 12*(2), 63–90.

Datta Gupta, N., Poulsen, A., & Villeval, M.-C. (2005). *Male and female competitive behavior: Experimental evidence* (IZA Discussion Paper No. 1833). Bonn, Germany: Institute for the Study of Labor.

Dimand, R. W., Forget, E. L., & Nyland, C. (2004). Retrospectives: Gender in classical economics. *Journal of Economic Perspectives, 18*(1), 229–240.

Doeringer, P., & Piore, M. J. (1971). *Internal labor markets and manpower adjustment.* New York: D. C. Heath.

Duguet, E., & Petit, P. (2006). Hiring discrimination in the French financial sector: An econometric analysis on field experiment data. *Annals of Economy and Statistics, 78,* 79–102.

England, P. (1982). The failure of human capital theory to explain occupational sex segregation. *Journal of Human Resources, 17*(3), 358–370.

England, P. (1992). *Comparable worth: Theories and evidence*. New York: Aldine de Gruyter.

Eurostat. (2009). The European Union labour force survey. Available at http://epp.eurostat.ec.europa.eu/portal/page/portal/labour_market/earnings/main_tables

Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica, 77*(3), 1637–1664.

Gneezy, U., Niederle, M., & Rustichini, A. (2003, August). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics,* 1049–1074.

Gneezy, U., & Rustichini, A. (2004, May). Gender and competition at a young age. *American Economic Review Paper and Proceedings,* 377–381.

Goldin, C. (2006). The rising (and then declining) significance of gender. In F. D. Blau, M. C. Brinton, & D. B. Grusky (Eds.), *The declining significance of gender?* New York: Russell Sage.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review, 90*(4), 715–741.

Gundersson, M. (1989). Male-female wage differentials and policy responses. *Journal of Economic Literature, 27*(1), 46–72.

Heckman, J. (1998). Detecting discrimination. *Journal of Economic Perspectives, 12,* 101–116.

Heckman, J., & Siegelman, P. (1993). The Urban Institute audit studies: Their methods and findings. In M. Fix & R. J. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 187–258). Washington, DC: Urban Institute Press.

Hellerstein, J., Neumark, D., & Troske, K. R. (2002). Market forces and sex discrimination. *Journal of Human Resources, 37*(2), 353–380.

Institute for Women's Policy Research. (2009). *The gender wage gap: 2008.* Available at http://www.iwpr.org/pdf/C350.pdf

Juhn, C., Murphy, K. M., & Pierce, B. (1991). Accounting for the slowdown in black-white wage convergence. In M. Kosters (Ed.), *Workers and their wages: Changing patterns in the United States* (pp. 107–143). Washington, DC: AEI Press.

Meng, X. (2004). Gender earning gap: The role of firm specific effects. *Labour Economics, 11*(5), 555–573.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy, 66*(4), 281–302.

Mincer, J., & Polachek, S. (1974). Family investments in human capital: Earnings of women. *Journal of Political Economy, 82*(2), S76–S108.

Nekby, L., Skogman Thoursie, P., & Vahtrik, L. (2007). Gender and self-selection into a competitive environment: Are women more overconfident than men? *Economics Letters, 100,* 405–407.

Neumark, D. M. (1996). Sex discrimination in restaurant hiring: An audit study. *Quarterly Journal of Economics, 111*(3), 915–941.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics, 122*(3), 1067–1101.

Oaxaca, R. L., & Ransom, M. L. (1999). Identification in detailed wage decompositions. *Review of Economics and Statistics, 81,* 154–157.

Olivetti, C., & Petrongolo, B. (2008). Unequal pay or unequal employment? A cross-country analysis of gender gaps. *Journal of Labor Economics, 26*(4), 621–654.

Pager, D. (2007). The use of field experiments for studies of employment discrimination: Contributions, critiques and directions for the future. *ANNALS of the American Academy of Political and Social Science, 609,* 104.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review, 62,* 659–661.

Polacheck, S. W. (1975). Discontinuous labor force participation and its effect on women's market earnings. In C. B. Lloyd (Ed.), *Sex, discrimination, and the division of labor* (pp. 90–124). New York: Columbia University Press.

Polachek, S. W. (1984). Women in the economy: Perspectives on gender inequality. In U.S. Civil Rights Commission (Ed.), *Comparable worth: Issue for the 80s* (Vol. 1, pp. 34–53). Washington, DC: Government Printing Office.

Reskin, B. F., & Bielby, D. D. (2005). A sociological perspective on gender and career outcomes. *Journal of Economic Perspectives, 19*(1), 71–86.

Riach, P. A., & Rich, J. (1987). Testing for sexual discrimination in the labour market. *Australian Economic Papers, 26,* 165–178.

Riach, P. A., & Rich, J. (2002). Field experiments of discrimination in the market place. *Economic Journal, 112,* F480–F518.

Riach, P. A., & Rich, J. (2006). An experimental investigation of sexual discrimination in hiring in the English labor market. *B.E. Journal of Economic Analysis & Policy, 6*(2), Article 1. Available at http://www.bepress.com/bejeap/advances/vol6/iss2/art1

Säve-Söderbergh, J. (2007). *Are women asking for low wages? An empirical analysis of individual wage bargaining and ability signalling* (Working Paper No. 7/2007). Stockholm: Swedish Institute for Social Research.

Small, D., Gelfand, M., Babcock, L., & Gettman, H. (2007). Who goes to the bargaining table? *Journal of Personality and Social Psychology, 93,* 600–613.

Weiselbaumer, D. (2004, Spring). Is it sex or personality? The impact of sex-stereotypes on discrimination in applicant selection. *Eastern Economic Journal, 30,* 159–186.

# 55

# Economics and Race

Sean E. Mulholland
*Stonehill College*

Race is a construction of society. These small, observable differences, often in pigmentation and physical characteristics, have, throughout time, played a role in defining and reinforcing other less easily observed differences. This is true both across groups and within groups. However, even using the term *groups* may be misleading. This spectrum of observable differences has two important consequences. First, unique individuals that may have little else in common are grouped as similar by others. Second, by being treated as a member of said group, one's own identity of self is altered. This chapter first looks at socioeconomic measures for those defined as members of certain racial groups. Traditional measures such as earnings, education, and segregation present the reader with foundational knowledge of racial differences across time and space. Second, the chapter discusses theories and empirical research that seek to explain these differences across and within these various racial groups, as well as policies seeking to reduce the disparate results.

To analyze the nexus between race and economics, a definition of race is required. Yet defining race as distinct, mutually exclusive categories is misleading at best. Physical characteristics, such as skin color, nose width, and hair texture, are often used to define race. Unfortunately, as any introductory anthropology text will tell you, human characteristics are on a continuous gradient, not in unique categories. Thus, much of the historical data presented and many of the academic works cited use basic categories such as black, white, and Asian in the analysis. New data and innovative researchers have begun to look within these groups, such as Asians with varying types of eyelids and the darkness of skin tone of blacks and African Americans, to gain a better grasp of the importance these subtle differences play within each traditional race category.

Attempting to group individuals into various categories in order to compare across and within groups often requires the use of averages or medians. No one is the average member of a group, and only one in millions is the median of that group. Thus, the average or median of the group one identifies with in no way reflects one's individual outcome.

One of the difficulties associated with empirical analysis is that correlation does not imply causation. Even though one group may have lower earnings or higher unemployment, this does not necessarily mean that it is due to current or past discrimination. For instance, as demonstrated in this chapter, African Americans have lower annual earnings than Asian Americans. Although one possible explanation may be current discrimination, another plausible hypothesis is that Asian Americans have more years of schooling, more years of experience, or another aspect that may affect the group's productivity and, therefore, earnings.

Characteristics, such as educational attainment, that affect a person's earnings may have been subject to past policies. The history of the United States includes a vast number of examples where minority groups have been at a disadvantage compared to settlers who were predominantly white, of northern European origins, and Protestant in faith. Any person who did not match this description had a significantly lower standing in early American society. Blacks were enslaved and perceived as objects, not as human beings; Chinese and Japanese immigrants were abused and denied citizenship, even into the second and third generations; Catholics were often treated with suspicion; and Jews were persecuted. Even the whites of southern or eastern European descent did not enjoy the same privileges and opportunities as the whites of, say, English

or German origins. So while it may be the case that discrimination is not present in the current labor market, legacy effects, or past discrimination and social policies, may affect current economic outcomes. Thus, when analyzing differences in outcomes, researchers attempt to include as many current and past characteristics in their analyses as possible.

This chapter first lays out some current and historical statistics on earnings, education, geographic distribution, marital status, and occupational choices of African Americans and whites. It then discusses recent developments in the economic literature on how economists have attempted to explain these differences between African Americans and whites. Next, the chapter focuses on Asian Americans and research seeking to explain their economic welfare. Then this chapter discusses the effects of the Civil Rights Act of 1964 and affirmative action. Because the chapter is about race, it focuses on differences between African Americans, Asians, and whites. Ethnic differences, such as the differences between Cuban and Dominican, or Korean and Chinese, though important because of their unique histories and experiences, are discussed only when the emphasis is on differences within race categories.

## Data on African Americans

Figure 55.1 shows the median earnings of full-time, year-round male workers by race in the United States from 1955 to 2007 as reported by the Current Population Survey of the U.S. Census.

These data, reported in real 2007 dollars, show that full-time, year-round white workers received higher earnings than African American workers. In 1955, the median earnings for full-time white male workers were $29,618, while African American workers earned about 61% of whites' amount: $18,033. Over time, however, full-time African American workers have witnessed a decline in earnings disparity. By 1980, African American males earned about 70% of male white workers' earnings. By 2007, the ratio of African American male to white male worker was about 0.78. Although this chapter focuses on differences within the United States, it is important to stress that race also matters in most other countries. For example, African Canadian men earn 18% less than Canadian whites, and nonwhite immigrants in Britain earn less than similarly skilled white immigrants (Howland & Sakellariou, 1993; Stewart, 1983).

Figure 55.2 tells a slightly different story for full-time, year-round female African American and white workers.

In 1955, the median earnings for full-time white female workers were $19,339, while for African Americans, the year's earnings were about 51% of whites' amount: $9,933. Over time, however, full-time African American female workers have witnessed a greater decline in earnings disparity than their male counterparts. By 1980, African American females earned about 93% of female white workers' earnings. By 2007, the ratio of African American female to white female worker had fallen to about 0.86.

These differences are often attributed to racial discrimination. However, at least a portion of this difference may occur even if market participants are not prejudiced. One possible nonprejudicial difference may result from differences in education. Figure 55.3 shows the average years of schooling for African American and white workers in the United States from 1880 to 2000 (Turner, Tamura, & Mulholland, 2008).

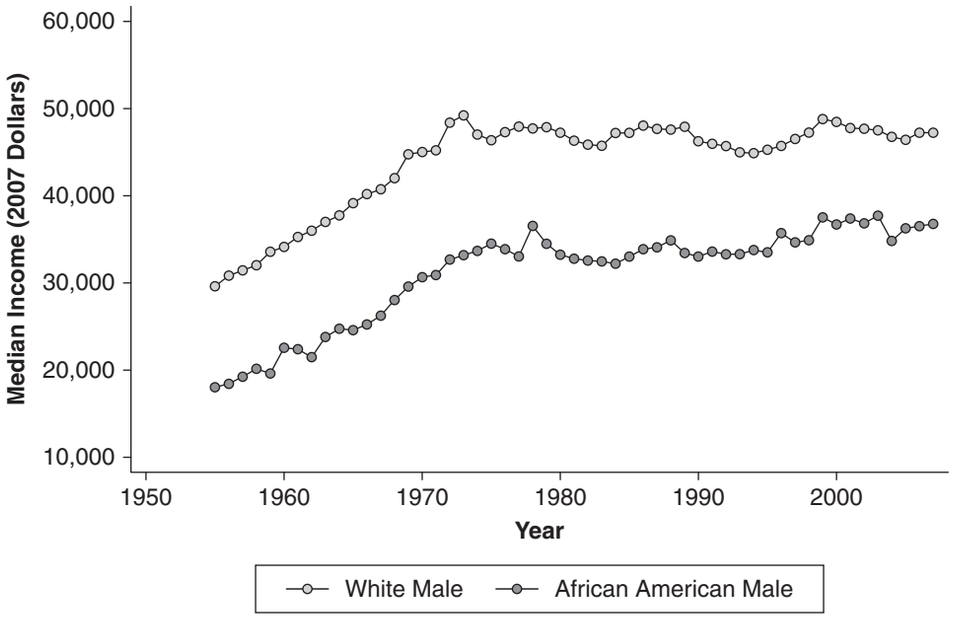


**Figure 55.1** Median Income: African American and White Male Workers, Full-Time, Year-Round, 1955–2007

SOURCE: U.S. Census Bureau (2008b).

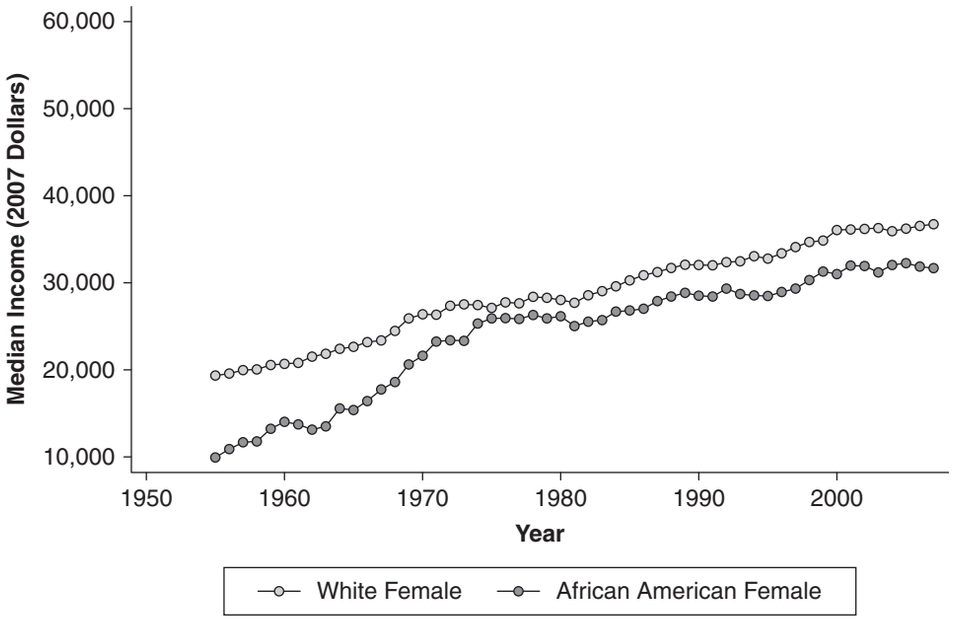


**Figure 55.2** Median Income: African American and White Female Workers, Full-Time, Year-Round, 1955–2007

SOURCE: U.S. Census Bureau (2008b).

Again, one sees a similar pattern: Whites averaged just over 4 years of schooling in 1880, while African Americans averaged just over one half of a year of schooling. Even with educational disenfranchisement among African Americans in the South, African Americans witnessed an increase in years of schooling. By 1950, the average years of schooling completed by an African American worker were just over 6.5 years. White workers in 1950 had 9.5 years of schooling. In 2000, African American workers possessed just over 12 years of schooling, while whites had completed just over 13 years of schooling. Thus, some of the differences in earnings may be explained by differences in educational attainment.

Furthermore, geography can play a role, especially if local labor markets differ. Table 55.1 shows the geographic distribution of whites and African Americans across the various regions of the United States. African Americans are located primarily in the South, while whites are more concentrated in the Northeast and West. John Bound and Richard Freeman (1992) emphasize that labor is largely

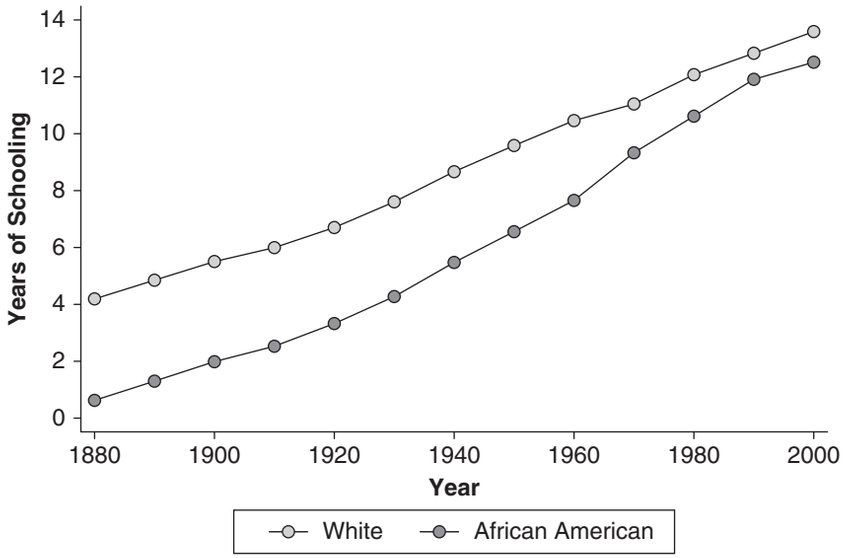


**Figure 55.3** Average Years of Schooling: African American and White Workers, 1880–2000

SOURCE: Turner, Tamura, and Mulholland (2008).

**Table 55.1** Geographic Distribution by Race: African Americans and Whites

| | *Percentage of Total Group Population* | |
|---|---|---|
| | *African American* | *White* |
| **Region** | | |
| Northeast | 18.00 | 19.54 |
| Midwest | 18.78 | 25.22 |
| South | 53.62 | 34.25 |
| West | 9.60 | 20.98 |

SOURCE: U.S. Census Bureau (2000).

NOTE: Percentages may not add to 100 due to rounding.

immobile in the short run, and these differences in regional location will also shape labor market outcomes, such as earnings. Because African Americans are more highly concentrated in the South, where wages and earnings are lower, a portion of the white–black earnings gap may be due to geographic location.

Another factor that affects the incentives and constraints of individuals is their family arrangements. Table 55.2 shows the marital status of African Americans and whites. Whites are much more likely to be married, with their spouse present, than African Americans. In addition, African Americans are much less likely to have ever married. Empirically, studies show anywhere from 0 to a 50% wage premium for married men. Theoretically, this marriage premium could arise from various sources. First, marriage can make men more productive by allowing them to specialize in nonhousehold production. Second, it could be that employers discriminate in favor of married men, and third, it could be that married men possess some unobservable characteristic that makes them more productive. Attempting to eliminate much of the unobservable characteristics, both Harry Krashinsky (2004) and Kate Antonovics and Robert Town (2004) use twin data to study the impact of marriage on wages. Although Krashinsky finds a 6% marriage premium for within-twin estimates, Antonovics and Town find a 27% married wage premium. Because African Americans are less likely to be married, they are either unable to specialize in nonhousehold production or less likely to possess the unobservable characteristic that makes them more productive.

The industry in which one chooses to work may also affect labor market outcomes. Table 55.3 shows the distribution of employment by industry and race. African Americans are more likely to work in industries that face more regulation: transportation and utilities, education and health, public administration, and the armed forces. African Americans are also more likely to work in information and leisure and hospitality. When one compares industrial distribution, the difficulty often arises as to whether occupation is a choice or a constraint. If one believes that firms in certain industries discriminate against certain types of employees, the data above may reflect more of a constraint and less of a decision by those employed.

**Table 55.2** Marital Status by Race: African Americans and Whites

| | *Percentage of Total Group Population* | |
|---|---|---|
| | *African American* | *White* |
| Married, spouse present | 31.8 | 53.9 |
| Married, spouse absent | 1.7 | 1.2 |
| Widowed | 6.2 | 6.1 |
| Divorced | 11.0 | 9.8 |
| Separated | 4.4 | 1.8 |
| Never married | 45.0 | 27.2 |

SOURCE: U.S. Census Bureau (2008a).

NOTE: Percentages may not add to 100 due to rounding.

Given that these differences may be choices not affected by race, economists must search for ways to isolate how and where discrimination is possible and, more important, how to disentangle its effects from these and other possible explanations.

## Theories of Discrimination

### Employer-Based Discrimination

In 1955, Gary S. Becker wrote his Nobel Prize–winning doctoral dissertation on the economics of discrimination. His work advances the theory that discrimination, contrary to the Marxist view, is costly to the individual who discriminates. He hypothesizes that because employer-based discrimination is costly, in the sense that employers must forgo profits in order to pay for their tastes for discrimination, increased competition will lower profits and thus the ability of discriminating firms' to practice discrimination. According to Becker (1957),

> If an individual has a "taste for discrimination," he must act as if he were willing to pay something, either directly or in the form of a reduced income, to be associated with some persons instead of others. When actual discrimination occurs, he must, in fact, either pay or forfeit income for this privilege. This simple way of looking at the matter gets at the essence of prejudice and discrimination. (p. 14)

**Table 55.3** Industrial Employment Distribution by Race: African Americans and Whites

| | *Percentage of Total Group Population* | |
|---|---|---|
| | *African American* | *White* |
| Agriculture, forestry, fishing, and hunting | 0.29 | 1.68 |
| Mining | 0.28 | 0.57 |
| Construction | 4.25 | 8.60 |
| Manufacturing | 9.88 | 11.09 |
| Wholesale and retail trade | 12.48 | 14.60 |
| Transportation and utilities | 7.78 | 5.02 |
| Information | 2.75 | 2.30 |
| Financial activities | 6.11 | 7.06 |
| Professional and business services | 9.61 | 10.92 |
| Educational and health services | 26.27 | 20.38 |
| Leisure and hospitality | 9.48 | 8.70 |
| Other services | 4.45 | 4.69 |
| Public administration | 6.31 | 4.38 |
| Armed forces | 0.04 | 0.01 |

SOURCE: U.S. Census Bureau (2008a).

NOTE: Percentages may not add to 100 due to rounding.

Therefore, the market structure in the output market affects the ability of individual firms to discriminate. By refusing to hire individuals based on characteristics that have nothing to do with productivity and thereby reducing the pool of possible employees, discriminating firms will, in the long run, face higher labor costs. Increased competition in the output market will force these firms to find lower cost production methods. Increased competition will squeeze higher cost firms, including discriminating firms, out of the market. According to Becker's theory, only nondiscriminating firms will continue to operate. The result also suggests an important implication for wage equality across races. In the long run, each employee will be paid his or her marginal product multiplied by the price of the output she or he is producing. Or stated another way, workers will be paid a wage that reflects exactly the value they add to the firm. Without any racial or other type of discrimination, wage gaps between equally skilled and productive whites and nonwhites should no longer exist.

This does not mean that average wage inequality across races will disappear or even decline. Kenneth Arrow (1972, 1973) and Edmund Phelps (1972) stress that a skills gap and imperfect information can explain racial wage differentials. African American workers have on average fewer years of schooling; thus, their wages may be lower. One way to test Becker's (1957) model is to analyze firms that have experienced an increase in competition through an exogenous change in regulation. In the 1970s, the banking system experienced such a regulatory change.

Before the 1970s, banks were unable to operate across state lines. Banking regulations up to this point required banks to operate as independent entities within each state border. Although Bank of America could have branches in any state, each state operation had to operate independently. Customers who wanted to change the address of their accounts were required to open new accounts if they moved across state lines. These restrictions limited banking competition to within the state borders; banks in one state were not in competition with banks in another.

Innovations, such as the automated teller machine, that make distance banking easier weaken the desire and ability of banks to fight regulatory changes that eliminated these state controls. Most states deregulated geographic restrictions on banking between the mid-1970s and 1994, when the federal Riegle-Neal Act effectively eliminated these restrictions. Granting banks the ability to compete across state borders increased the level of competition in the banking industry.

Becker's theory predicts that such an increase in competition will lead to the elimination or, at least, the reduction of a wage gap that exists due to racial discrimination. To test the validity of that prediction, Ross Levine, Alexey Levkov, and Yona Rubinstein (2008) collected data on earnings of black and white non-Hispanic male workers aged between 18 and 65 from 1977 to 2007. They also compared the actual ratio of interracial marriages in each state from the 1970 census to the same ratio based on hypothetical random pairing of partners to assess the level of racial bias. Their research showed that in the states with high levels of racial prejudice, the increase in market competition significantly reduced the wage gap between African American and white workers. In the states with lower-than-average bias, where the actual and hypothetical interracial marriage ratios were very similar, the reduction was almost unnoticeable. Their results suggest that Becker's theory holds empirically only when the initial degree of racial bias in the economy is high. If the degree of racial bias is already low, less of the wage differential is due to racial discrimination, and thus, increases in competition will have little effect on racially based wage differentials.

The banking industry is not the only example of increased competition in the output market due to regulatory changes. Looking at the racial wage gap in the for-hire

trucking industry, Nancy Rose (1987) finds that increased competition due to deregulation reduced the racial wage gap in the trucking industry. Expanding on this work, James Peoples and Lisa Saunders (1993) find evidence that deregulation reduced the wage gap for both union and nonunion drivers. Peoples and Rhoda Robinson (1996) extend this line of research to the telecommunications industry by showing that the earning disparity between black and white males decreased with the divestiture of AT&T and the resulting increase in competition. John Heywood (1998) examines the effects of deregulation in airlines, trucking, rail, and telecommunications on racial earnings. His results suggest that all industries, except airlines, witnessed a decline in the racial earnings gap.

Becker (1957) writes that, in the long run, discriminating businesses, operating in a competitive market, will become less profitable than the nondiscriminating ones, lose market share, and eventually fail. In response, discriminating firm owners could simply sell their firms, offer their labor for a salary, and earn higher returns. However, rational people seek to maximize their utility, not their monetary earnings alone. Though monetary income has a large effect on utility, prejudiced employers would rather suffer pecuniary losses than employ minority workers. Kerwin Charles and Jonathan Guryan (2007) claim that, because of such preferences, those formerly prejudiced employers would remain prejudiced as employees and work only at firms whose owners were equally prejudiced. In an economy with a significant number of such white, racially biased workers, there might still be some prejudiced firms at which these biased employees will work. If at least one such firm operated, wage gaps could continue to exist.

In the same study, Charles and Guryan (2007) also find that there is a close relationship between the level of prejudice and the magnitude of the wage gap. They construct a prejudice measure by using six questions asked in the General Social Survey (GSS) from 1972 to 2004. These questions included respondents' "feelings about interracial marriage, their sense of whether racially restrictive housing covenants were appropriate, their views about children being racially segregated in schools, and their view on whether the government should be obligated to help blacks." They find the racial wage gap to be larger in communities with higher levels of prejudice. Charles and Guryan conclude that the most prejudiced region in the United States is the Southeast, while New England and the West Coast are the least.

## Customer-Based Discrimination

Though Becker's model predicts that increased competition will eliminate employment discrimination in the long run, this is not true for customer discrimination. Customer discrimination exists when customers have and are willing to pay for their discriminatory preferences. If customers wish to avoid contact with a certain group of people, customers will repeatedly pay more to avoid contact with individuals from that group. Lawrence Kahn and Derek Shearer (1988) investigate racial differences in the 1985 to 1986 salaries of individual basketball players and the attendance effects of replacing one black player with a white player. Holding constant various factors, such as productivity, market, and player draft position, they find that black players earn 20% less than white players. In addition, they show that "replacing one black player with an identical white player raises home attendance by 8,000 to 13,000 fans per season" (p. 40).

Even without direct contact, customer discrimination may be present. Clark Nardinelli and Curtis Simon (1990) address this possibility by analyzing the value as listed in the 1989 issue of *Beckett Baseball Card Price Guide* of Topps baseball cards issued in 1970. Because the two important measures of card value, player's lifetime performance and the number of cards issued, are easy to measure, they are able to construct a model to determine the effect of race on card value. They find a card value gap of 10% among hitters of comparable ability and a 13% gap in card value among pitchers. Torben Anderson and Sumner La Croix (1991) look at baseball cards from 1977, when the number of cards issued did not vary by player, and find results that support those found by Nardinelli and Simon.

Keith Ihlanfeldt and Madelyn Young's (1994) work on fast-food restaurants in Atlanta show that a 10-percentage-point increase in white customers reduces African American wages by about 1%. Ihlanfeldt and David Sjoquist (1991) use the racial composition of residents in subcounty areas as proxies for customer composition and find that the racial composition of an employment area affects the type of job held by black workers. Using data from four metropolitan areas, Holzer and Ihlanfeldt (1998) show that racial composition of the customers affects whom employers hire. They find that the magnitude of these effects vary by "occupational category and the degree of direct contact with customers on the job" (p. 862). Holzer and Ihlanfeldt suggest that these differences may be increasing in importance as the distribution of jobs shift from those with little customer contact, such as manufacturing, to those with greater customer contact, such as retail.

## Neighborhood Effects

Customer discrimination may also account for the failure of inner-city blacks to migrate to the suburbs (Kain, 1968). Transportation difficulties or information limitations may also result in spatial mismatches (Holzer, Ihlanfeldt, & Sjoquist, 1994). Unfortunately, discrimination in the credit market may reduce the ability of blacks to move households or businesses. Faith Ando (1988) finds blacks are less likely to be approved for loans even after accounting for various factors that influence approval rates. David Blanchflower, Phillip Levine, and David Zimmerman (2003) find that "black-owned firms, in particular, are substantially more likely to be denied credit than other groups and are charged higher interest rates for those loans that are approved than

are other firms that are otherwise comparable" (p. 930). Furthermore, a study of the Boston mortgage market during the early 1990s reveals substantial racial differences in the likelihood that a mortgage application would be rejected, even after controlling exhaustively for differences in credit history and various other factors (Munnell, Tootell, Browne, & McEneaney, 1996).

The lack of migratory ability leads to another possible reason for differences in earnings discussed in a vast literature on peer group effects or social interaction and neighborhood effects (Borjas, 1995; Case & Katz, 1991; Glaeser, Sacerdote, & Scheinkman, 1996). Building on this work and the work of Kain (1968), David Cutler and Edward Glaeser (1997) find African Americans in segregated communities are worse off than those in nonsegregated communities. Cutler and Glaeser show that a one-standard-deviation reduction in segregation would eliminate one third of the education and earnings gap between whites and blacks.

## Statistical Discrimination

Statistical discrimination, first described by Arrow (1972, 1973) and Phelps (1972) and further developed theoretically by Joseph Altonji and Charles Pierret (2001), occurs when an individual's estimated productivity is based on the group to which he or she is categorized and not by his or her individual characteristics. William Wilson (1996) reveals that both black and white employers are reluctant to hire young, black urban males. In addition, "statistical discrimination can be quite damaging to both the efficiency of market allocations and to equity. This is due to the very real possibility that the empirically valid statistical generalizations lying at the heart of such discrimination can be self-fulfilling prophecies" (Loury, 1998, p. 123; see also Coate & Loury, 1993; Lundberg & Startz, 1983).

## Recent Research on African American Discrimination

These categorizations can be based on a large range of factors. Racial stereotypes that are the bases of prejudices are often identified with one or more physical characteristics. But as suggested theoretically by Kevin Lang (1986), speech patterns and nonverbal communication may also be a source of discrimination. A recent study of speech patterns carried out by Jeffrey Grogger (2008) collected the phone recordings from interviews for the National Longitudinal Survey of Youth and removed all the information that could reveal each interviewee's gender, race, and age. He then asked people to identify the speaker of each recording as either black or white. Comparing the reported earnings of the black interviewees identified as sounding black to those black interviewees identified as sounding white, he finds that blacks who sound black earn 10% less than blacks who do not. In comparison, whites who sound black earn 6% lower salaries than whites who do not.

Not only does it appear that African American–sounding speech may be associated with wage differentials, but a person's name may also play a role. Marianne Bertrand and Sendhil Mullainathan (2003) test whether having a typically white-sounding name is more advantageous for job applicants than having a black-sounding name. Using the frequency of names given to African American and white newborns in the state of Massachusetts between 1974 and 1979, they sent four resumes to each job posting in *The Boston Globe* and the *Chicago Tribune.* Each group of four resumes, which included detailed information, had two of high quality and two of low quality. One high-quality and one low-quality resume were always given distinctly African American names, while the other two would always get white-sounding names.

Resumes with white-sounding names had a callback rate of 10%, while black-sounding names had a rate of only 6.7%. The results suggest that an applicant with a white-sounding name can, on average, expect an invitation for a job interview for every 10 resumes submitted, while an applicant with an African American–sounding name will only receive 1 for every 15. This result suggests that employer-based discrimination plays a role in the interview process.

However, statistical discrimination and neighborhood effects may also be present. Bertrand and Mullainathan's (2003) experiment also looks at the effect a person's address has on whether an applicant receives an invitation for an interview. Consistent with neighborhood effects, applicants whose neighborhood residence was whiter, with higher earnings and levels of education, realized a greater callback rate than those whose address represented the lower earning, less educated, predominately African American neighborhoods. However, living in a whiter, higher earning, more educated area neighborhood does not help the applicants with distinctively black names very much.

Moreover, employers seemed to be more willing to acknowledge the difference between high- and low-quality white applicants than differences between high- and low-quality African American applicants. High-quality white applicants received invitations for interviews from about 11% of their applications, while low-quality white applicants received invitations from only 8.8%. In contrast, high-quality blacks experienced a callback rate of just less than 7%, while low-quality blacks received callbacks on 6.4%.

Speculating that distinctively African American–sounding names convey information about socioeconomic background, Roland Fryer and Steven Levitt (2004) seek to determine consequences and causes of black-sounding names. Using California birth certificates, they collected data on the first names given to black and white non-Hispanic newborns between 1961 and 2000. Using these data, they created the black name index (BNI) to measure the distinctiveness of a given name by dividing the fraction of black children given that name over the fraction of both black and white children receiving the exact name. Babies born in predominantly black hospitals, over time, were more likely to receive distinctively black names than those born in predominantly white hospitals. Blacks living in segregated

neighborhoods were more likely to give their children distinctively black names than those living in integrated communities. Moreover, a woman's first name can indicate to her potential employer the circumstances she grew up in and her current situation, both of which can imply her labor productivity. Fryer and Levitt also found that a woman with a BNI of 100 (a name that no whites have) is 20.9% more likely to have been born to a teenage mother and 31.3% more likely to have been born out of wedlock than a similar black woman with a BNI of around 50. Once a potential employer realizes that a resume has been submitted by an individual with a distinctively black name, he or she has a reason to believe the applicant has poor credentials without reading it in detail. This could explain why black applicants receive so few callbacks, regardless of their qualifications. If this is employer-based discrimination, instead of statistical discrimination, then the employers simply reject the applicant because of their taste for discrimination.

Focus on African Americans versus white, however, may miss a subtle yet important difference in how various shades of discrimination are expressed. Preferences for those with lighter pigmentation have been cited by historians who show that both whites and lighter skinned blacks often found ways to maintain distinct social networks (Gatewood, 2000). These culturally defined restrictions have serious consequences. Robert Margo (1992) finds that light-complected slaves were more likely to receive skill training. Shawn Cole (2005) reports that 40% of freed slaves were mulattoes, and less than 6% of slave sales in late eighteenth- and early nineteenth-century Louisiana involved mulatto slaves.

Looking at Los Angeles neighborhoods in the 1900s, James Johnson, Elisa Bienenstock, and Jennifer Stoloff (1995) discover that the combination of black racial identity and a dark skin tone reduces an individual's odds of working by 52%, after controlling for education and including quality, age, and criminal record. A lighter complexion is also associated with higher incomes and life chances than a darker complexion in blacks in the United States (Keith & Herring, 1991; Ransford, 1970). Andrew Goldsmith, Darrick Hamilton, and William Darity Jr. (2006) find that blacks with lighter-colored skin earn higher wages than black workers whose skin color is identified as medium and dark. The wage gap between medium-skinned blacks and whites is 16%; the wage gap between dark-skinned blacks and whites is 17%. In comparison, the gap between light-skinned blacks and whites is only 7.6%.

## Data on Asian Americans

Repeating the data exercise performed with African Americans but substituting data on Asian Americans reveals a different story. Even though it has been fewer than 70 years since some of their ancestors were held in internment camps, Asian Americans seem to be faring better than the population at large.

These data, reported in real 2007 dollars, show that full-time, year-round white and Asian American male workers have, at least for the last 20 years, similar earnings.

In 1988, the median earnings for full-time white male workers was $47,598, while for Asian Americans, the yearly earnings were about 99% of whites' earnings: $47,287. Over time, however, full-time Asian American male workers have witnessed slightly higher earnings growth than white male workers. By 2007, Asian American males earned about 108% of what white male workers earned.

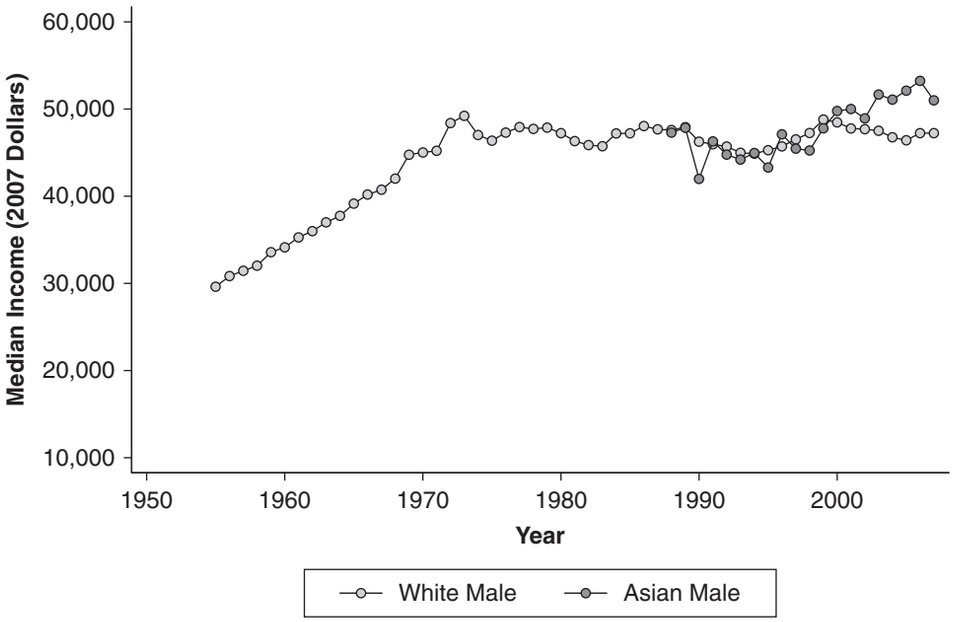


**Figure 55.4** Median Income: Asian American and White Male Workers, Full-Time, Year-Round, 1955–2007

SOURCE: U.S. Census Bureau (2008b).

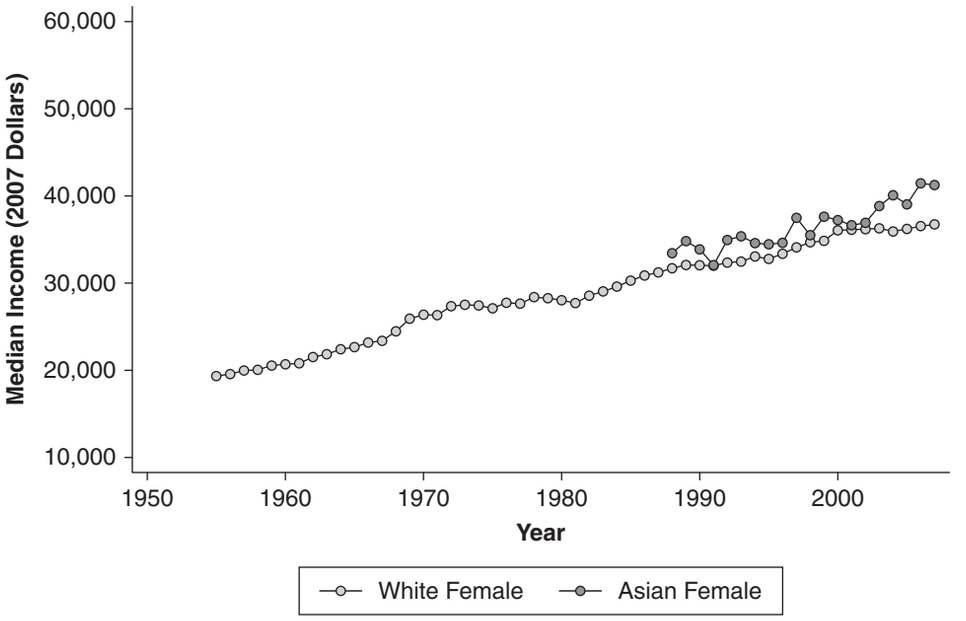


**Figure 55.5** Median Income: Asian American and White Female Workers, Full-Time, Year-Round, 1955–2007

SOURCE: U.S. Census Bureau (2008b).

Figure 55.5 shows a similar story for full-time, year-round female Asian American workers relative to their white female counterparts.

In 1988, the median income for full-time white female workers was $31,701, while Asian American women earned about $33,426. By 2000, Asian American females earned about 104% of what white female workers earned. By 2007, the ratio of Asian American female earnings to white female worker earnings was about 1.12.

Presenting the difference in education will help to suggest why both Asian American females and males now earn more than their respective white colleagues. Figure 55.6 shows the average years of schooling for Asian American and white workers in the United States from 1940 to 2000 (Ruggles et al., 2008)

In the 1940s, whites had just over 7 years of schooling, while Asian Americans had just fewer than 6.9 years of schooling. By 1980, Asian Americans had on average

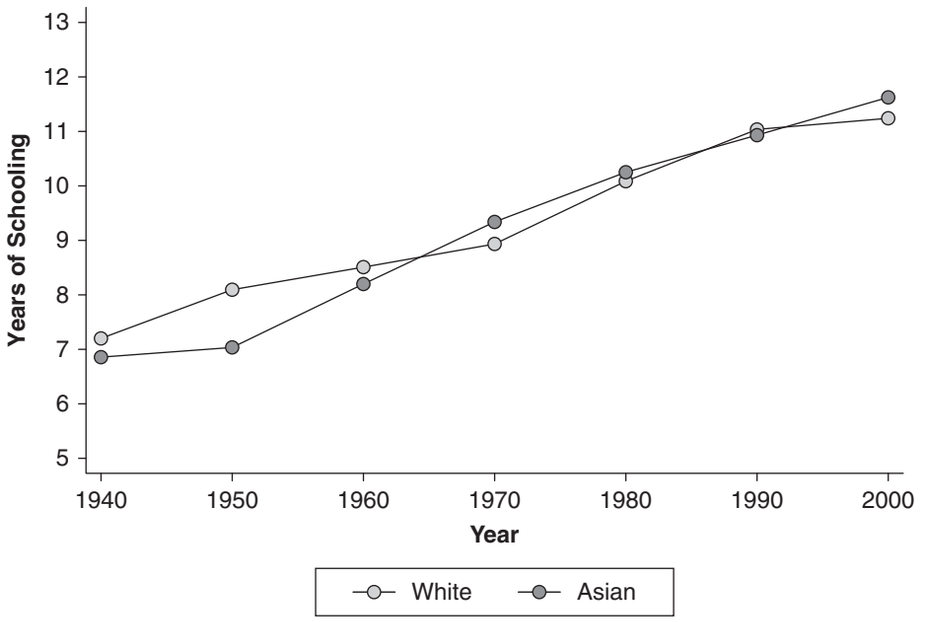


**Figure 55.6** Average Years of Schooling: Asian American and White Workers, 1940–2000

SOURCE: Data from Ruggles et al. (2008).

obtained more years of schooling than whites. The ratio of the average years of schooling for Asian Americans relative to whites was 1.04. From 1980 to 2000, this ratio varied little.

Geography can also play a role, especially if local labor markets differ. Table 55.4 shows the geographic distribution of whites and Asian Americans across the various regions of the United States. Asian Americans are primarily located in the west and northeast, much like their fellow white citizens, where wages and earnings are higher.

Another factor that affects the incentives and constraints of individuals is their family arrangement. Table 55.5 shows the marriage status of Asian Americans and whites. As presented above, empirically, studies show anywhere from 0 to a 50% wage premium for married men. Because Asian Americans are more likely to be married and less likely to be divorced than white Americans, Asian Americans can more easily specialize in nonhousehold production, or they are more likely to possess the unobservable characteristic that makes them more productive. Asian Americans are also less likely to be divorced than whites.

Industry of choice may also affect labor market outcomes. Table 55.6 shows the distribution of employment by industry and race. Although Asian Americans are more likely than whites to work in some industries that face greater regulation, such as education and health, they are also more likely to work in industries that face much less regulation: manufacturing, information, financial, professional and business services, leisure and hospitality, and other services. As mentioned above, the difficulty often arises as to whether occupation is a choice or a constraint. If one believes that firms in certain industries discriminate against certain types of employees, the data above may reflect a constraint and not a decision by those employed.

## Research on Asian American Discrimination

The data presented in Table 55.6 suggest that Asian Americans have attained higher levels of earnings, education, and family stability than the overall population, the other minorities, or both. A study conducted by Barry Chiswick in 1983 seems to reinforce the idea that Asian Americans have achieved higher socioeconomic welfare. Using data from the 1970 census of population, which identifies Asian Americans by their origins—Chinese, Japanese, and Filipino—Chiswick compares U.S.–born Asian male workers' earnings and employment to a control sample of U.S.–born whites. Chiswick finds that Asian American men had higher levels of earnings, employment, and education in 1969 than white men. In aggregate, the Asian Americans earned about $10,000, while white men earned only $9,900; Asian Americans had completed more schooling (12.6 vs. 11.9 years), and worked more hours (48.7 vs. 48.3 weeks). However, only 77% of Asian American men were married and living with their wives, compared to a rate of 86% for white men.

Within the Asian American group, the Chinese Americans and Japanese Americans performed better in all categories than the whites; however, the Filipino Americans fared worse. In 1969, Chinese Americans earned $10,400, had completed 13.1 years of schooling, and worked 47.6 weeks; the Japanese Americans earned $10,300, had been in school for 12.7 years, and worked 49.4 weeks. The Filipino Americans earned only $7,000, had completed 11.3 years of schooling, and worked 46.8 weeks. After controlling for human capital, demographics, and geographic area variables, the weekly earnings for those of Chinese, Japanese, and Filipino origins were lower than those of the whites by 2%, 4%, and 16%, respectively. But because both Chinese Americans and Japanese

**Table 55.4** Geographic Distribution by Race: Asian Americans and Whites

| | *Percentage of Total Group Population* | |
|---|---|---|
| | *Asian American* | *White* |
| Region | | |
| Northeast | 15.03 | 19.54 |
| Midwest | 11.80 | 25.22 |
| South | 20.82 | 34.25 |
| West | 52.35 | 20.98 |

SOURCE: U.S. Census Bureau (2000).

NOTE: Percentages may not add to 100 due to rounding.

**Table 55.5** Marital Status by Race: Asian Americans and Whites

| | *Percentage of Total Group Population* | |
|---|---|---|
| | *Asian American* | *White* |
| Married, spouse present | 58.3 | 53.9 |
| Married, spouse absent | 2.6 | 1.2 |
| Widowed | 4.5 | 6.1 |
| Divorced | 4.3 | 9.8 |
| Separated | 1.4 | 1.8 |
| Never married | 28.9 | 27.2 |

SOURCE: U.S. Census Bureau (2008a).

NOTE: Percentages may not add to 100 due to rounding.

Americans worked more weeks than the whites, their annual earnings were higher. On the other hand, the Filipino Americans had both significantly lower weekly earnings and more workable hours. Chiswick (1983) concludes that those of Chinese and Japanese origin, despite being minorities, are as successful in the labor market as whites and therefore do not suffer any discrimination.

Pramod Junakar, Satya Paul, and Wahida Yasmeen (2004) claim that the results obtained by Chiswick (1983) might not be very accurate because "estimates of discrimination based on studies of earnings are likely to be underestimates as they usually ignore the probability of finding employment in the first place" (p. 6). However, the year analyzed by Chiswick, 1969, was a year of full employment, with the unemployment rate of only 2.1%—the lowest in the post–World War II period. Therefore, the probability of being without a job was very small, regardless of race or ethnicity.

Even though Chiswick (1983) finds no evidence of discrimination against the Chinese and Japanese born in the United States and some researchers have subsequently concluded similarly, others have found that Asian Americans still earn less than similarly educated whites. One explanation for this discrepancy is the heterogeneity within Asian Americans. Although Chinese Americans and Japanese Americans tend to do well with respect to education and earnings, other nationalities, such as the Filipino Americans and southeastern Asians do not fare as well. The combination of these Asian subgroups into one large category creates the gap between the Asian Americans in general and whites in data from the 1970s.

Another possible weakness of Chiswick's (1983) study is the exclusion of foreign-born Asian Americans. A study conducted by Zhen Zeng and Yu Xie (2004) looks at place of birth and education to explain why there is no consensus on whether the Asian Americans have lower or higher earnings than whites and, if they do, why this gap in earnings exists between similarly educated Asian Americans and whites. Using 1990 census data supplemented with data from the National Survey of College Graduates, they divide Asian American male workers into three distinct groups: U.S.–born Asian Americans (UBA), U.S.–educated Asian immigrants (UEAI), and foreign-educated Asian immigrants (FEAI). FEAI are compared to the UEAI to assess the effects of the place of education on earnings, UEAI to UBA for the effects of nativity, and UBA to U.S.–born whites (UBW) for the effects of race.

Zeng and Xie (2004) show that UEAIs have 2 more years of schooling and 37% higher earnings than FEAI. In addition, one third of UEAIs work in professional occupations, while only 13% of FEAIs do. Although Asian Americans have higher earnings overall than whites, they earn less at each level of education, which means they have to obtain more education in order to have parity in earnings. UEAIs earn more than FEAIs but still less than UBAs. That said, it seems as though earnings differ both by nativity and by place of education. The results from the two models show that UBWs have the highest earnings, followed UBAs (5% less than UBW), UEAIs (5% less than UBA), and FEAIs (16% less than UEAI). However, only the difference between UEAI and FEAI is statistically significant. Although the effects of race and nativity are very small and negligible, where Asian Americans are educated has a substantial effect on their earnings. The results show that much of the discrimination of Asian Americans is directed against the immigrants, especially those who have completed their education prior to arriving in the United States. This suggests that Chiswick's (1983) decision to only include U.S.–born Asian Americans may be driving his results.

Chiswick's (1983) results might also be biased because he includes wages, salaries, and earnings from self-employment. Using data from the 1992 Characteristics of Business Owners (CBO) survey, Alicia Robb and Robert Fairlie (2007) show that Asian American–owned firms are 20.6% more likely to have profits of $10,000 or more,

**Table 55.6** Industrial Employment Distribution by Race: Asian Americans and Whites

| | *Percentage of Total Group Population* | |
|---|---|---|
| | *Asian American* | *White* |
| Agriculture, forestry, fishing, and hunting | 0.38 | 1.68 |
| Mining | 0.07 | 0.57 |
| Construction | 2.89 | 8.60 |
| Manufacturing | 13.44 | 11.09 |
| Wholesale and retail trade | 12.28 | 14.60 |
| Transportation and utilities | 3.92 | 5.02 |
| Information | 2.33 | 2.30 |
| Financial activities | 7.52 | 7.06 |
| Professional and business services | 13.97 | 10.92 |
| Educational and health services | 21.48 | 20.38 |
| Leisure and hospitality | 12.08 | 8.70 |
| Other services | 6.20 | 4.69 |
| Public administration | 3.45 | 4.38 |
| Armed forces | 0.00 | 0.01 |

SOURCE: U.S. Census Bureau (2008a).

NOTE: Percentages may not add to 100 due to rounding.

27.2% more likely to have employees, and have 60% higher sales than white-owned businesses. Asian American–owned firms also have higher survival rates: an 18% probability of closure versus 23% for white-owned firms. By including self-employed Asian Americans, who on average are much more successful than the typical owner, Chiswick's results might be upwardly biased. Moreover, people who are self-employed usually do not face the kind of discrimination from their employers that wage and salary earners face; thus, including business owners in an attempt to measure employer discrimination may reduce the likelihood of findings even if customer and statistical discrimination is present.

Because the growth of the Asian population in the United States is a recent phenomenon, there are relatively few studies that analyze the labor market outcome experienced by Asian Americans. The few studies available seem to disagree as to whether Asians in the United States are subject to discrimination in the labor market. The problems that scholars have to confront are lack of adequate data and the heterogeneity of the Asian group in terms of their ethnicity, nativity, and education.

## Policy

Two policies most often cited as having the greatest impact on racial discrimination are the Civil Rights Act of 1964 and affirmative action. Before the enactment of the Civil Rights Act of 1964, employment and contractual actions based on race were legal. The act established the Equal Employment Opportunity Commission (EEOC) and prohibits employment discrimination by large employers (more than 15 employees), whether or not they have government contracts. The resulting reduction in the black–white wage gap was large and abrupt (Card & Krueger, 1992; Donohue & Heckman, 1991).

Though the Civil Rights Act of 1964 sought to prohibit current discrimination, affirmative action was developed to overcome persistent disparate outcomes. The result was a system of preferences for minority college applicants, minority job applicants, and minority-owned businesses. Affirmative action has generated much debate on both constitutional and equity grounds. Although space here does not allow for a thorough discussion of the challenges to current court decisions or a thorough presentation of research looking at the redistributive and efficiency effects of affirmative action, a few noted theoretical and empirical studies are presented.

Finis Welch (1976) develops a theoretical model with taste-based discrimination and a quota-based system to show that affirmative action may cause unskilled workers to be assigned to skilled positions and vice versa. Lundberg and Startz (1983) develop a model to show that affirmative action can increase efficiency in human capital investment but lower efficiency in the labor market, resulting in an ambiguous net effect. Coate and Loury (1993) find that affirmative action can, theoretically, have negative effects on those it seeks to help. They show that if "identifiable groups are equally endowed ex ante, affirmative action can bring about a situation in which employers (correctly) perceive the groups to be unequally productive, ex post" (p. 1220).

Empirically, Jonathan Leonard (1984a) finds no negative productivity effects on federal contractors after affirmative action implementation. Leonard (1984b) finds that the share of employment accounted for by minorities rose at contractor establishments covered by affirmative action between 1974 and 1980, while those accounted for by white males declined. Using microlevel data, Holzer and David Neumark (2000b) discover that employers using affirmative action in job searches recruit "more extensively and screen more intensively" and "are more likely to ignore stigmatizing personal characteristics and histories when they hire" (p. 269). These affirmative action hires, however, come at the expense of white males, who see a 10% to a 15% decline in employment at these establishments.

In a meta-analysis of sorts, Stephan Thernstrom and Abigail Thernstrom (1997) report that black enrollments as a percentage of all enrollments in schools, excluding historically black colleges and universities, rose from 1.8% in 1960 to 9.0% in 1994. William Bowen and Derek Bok (1998) report that from 1960 to 1995 the percentage of blacks aged 25 to 29 who had graduated from college rose from 5.4% to 15.4%. Looking at data for selective institutions, Frederick Vars and Bowen (1998) and Bowen and Bok (1998) find some evidence that test scores are worse predictors for blacks. Thus, because African Americans with similar standardized test scores to whites' perform better in college and are more likely to graduate than their white classmates, affirmative action may have simply reduced the bias present in the college application process.

Although the program may place slightly less qualified individuals into positions and possibly lead to negative fit or stigma effects, it may also create positive externalities. Minorities benefiting from affirmative action may generate social benefits, such as working in underserved areas, which do not generate private pecuniary benefits. Overall, Holzer and Neumark (2000a) conclude, "Affirmative action offers significant redistribution toward women and minorities, with relatively small efficiency consequences" (p. 559).

## Conclusion

Race continues to play a role in whether and how employees are hired and compensated, the level of education one receives, and where and how one lives. As has been shown, many of these effects can be measured in various manners. Competitive markets can, theoretically in the long run, reduce employer-based discrimination and eliminate earnings and wage differences based on characteristics that

have no effect on labor productivity. However, customer and statistical discrimination can persist. Unfortunately, discrepancies in employment opportunities, wages, neighborhood effects, and loan approval can simply occur because of one's name, accent, address, or physical characteristics. Even with equal opportunity legislation and (or despite) affirmative action, African Americans appear to experience a greater difference in earnings relative to whites and Asian Americans.

## References and Further Readings


Altonji, J. G., & Pierret, C. R. (2001). Employer learning and statistical discrimination. *Quarterly Journal of Economics, 116*(1), 313–350.

Anderson, T., & La Croix, S. J. (1991). Customer racial discrimination in Major League Baseball. *Economic Inquiry, 29,* 665–677.

Ando, F. (1988). Capital issues and the minority-owned business. *Review of Black Political Economy, 16*(4), 77–109.

Antonovics, K., & Town, R. (2004). Are all the good men married? Uncovering the sources of the marital wage premium. *American Economic Review, 94*(2), 317–321.

Arrow, K. (1972). Some mathematical models of race in the labor market. In A. H. Pascal (Ed.), *Racial discrimination in economic life* (pp. 187–204). Lexington, MA: Lexington Books.

Arrow, K. (1973). The theory of discrimination. In O. A. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton, NJ: Princeton University Press.

Becker, G. S. (1957). *The economics of discrimination.* Chicago: University of Chicago Press.

Bertrand, M., & Mullainathan, S. (2003, January). *Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination* (Poverty Action Lab Paper No. 3). Cambridge: MIT.

Blanchflower, D. G., Levine, P. B., & Zimmerman, D. J. (2003). Discrimination in the small business credit market. *Review of Economics and Statistics, 85*(4), 930–943.

Borjas, G. (1995). Ethnicity, neighborhoods, and human capital externalities. *American Economic Review, 85*(3), 365–390.

Bound, J., & Freeman, R. (1992). What went wrong? The erosion of the relative earnings and employment among young black men in the 1980s. *Quarterly Journal of Economics, 107,* 201–232.

Bowen, W. G., & Bok, D. (1998). *The shape of the river.* Princeton, NJ: Princeton University Press.

Card, D., & Krueger, A. B. (1992). School quality and black-white relative earnings: A direct assessment. *Quarterly Journal of Economics, 107*(1), 151–200.

Case, A. C., & Katz, L. F. (1991). *The company you keep: The effects of family and neighborhood on disadvantaged youths* (NBER Working Paper No. 3705). Cambridge, MA: National Bureau of Economic Research.

Charles, K. K., & Guryan, J. (2007). *Prejudice and the economics of discrimination* (NBER Working Paper No. W13661). Cambridge, MA: National Bureau of Economic Research.

Chiswick, B. R. (1983). An analysis of the earnings and employment of Asian-American men. *Journal of Labor Economics, 1*(2), 197–214.

Coate, S., & Loury, G. C. (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review, 83*(5), 1220–1240.

Cole, S. (2005). Capitalism and freedom: Manumissions and the slave market in Louisiana, 1725–1820. *Journal of Economic History, 65*(4), 1008–1027.

Cutler, D., & Glaeser, E. (1997). Are ghettos good or bad? *Quarterly Journal of Economics, 112*(3), 827–872.

Donohue, J. J., III, & Heckman, J. (1991). Continuous versus episodic change: The impact of civil rights policy on the economic status of blacks. *Journal of Economic Literature, 29*(4), 1603–1643.

Fryer, R., & Levitt, S. (2004). The causes and consequences of distinctively black names. *Quarterly Journal of Economics, 119* (3), 676–805.

Gatewood, W. B. (2000). *Aristocrats of color: The black elite, 1880–1920.* Fayetteville: University of Arkansas Press.

Glaeser, E. L., Sacerdote, B., & Scheinkman, J. A. (1996). Crime and social interactions. *Quarterly Journal of Economics, 111*(2), 507–548.

Goldsmith, A., Hamilton, D., & Darity, W., Jr. (2006). From dark to light: Skin color and wages among African Americans. *Journal of Human Resources, 42*(4), 701–738.

Grogger, J. (2008, July). *Speech patterns and racial wage inequality* (Harris School Working Paper No. 08.13). Chicago: University of Chicago.

Heywood, J. S. (1998). Regulated industries and measures of earnings discrimination. In J. Peoples (Ed.), *Regulatory reform and labor markets* (pp. 287–324). Boston: Kluwer Academic.

Holzer, H. J., & Ihlanfeldt, K. (1998). Customer discrimination and the employment outcomes of minorities. *Quarterly Journal of Economics, 113*(3), 833–865.

Holzer, H. J., Ihlanfeldt, K. R., & Sjoquist, D. L. (1994, May). Work, search, and travel among white and black youth. *Journal of Urban Economics,* 320–345.

Holzer, H. J., & Neumark, D. (2000a). Assessing affirmative action. *Journal of Economic Literature, 38,* 483.

Holzer, H. J., & Neumark, D. (2000b). What does affirmative action do? *Industrial Labor Relations Review, 53*(2), 240–271.

Howland, J., & Sakellariou, C. (1993). Wage discrimination, occupational segregation and visible minorities in Canada. *Applied Economics, 25*(11), 1413–1422.

Ihlanfeldt, K. R., & Sjoquist, D. L. (1991). The role of space in determining the occupations of black and white workers. *Regional Science and Urban Economics, 21*(2), 295–315.

Ihlanfeldt, K. R., & Young, M. V. (1994). Intrametropolitan variation in wage rates: The case of Atlanta fast-food restaurant workers *Review of Economics and Statistics, 76*(3), 425–433.

Johnson, J. H., Jr., Bienenstock, E. J., & Stoloff, J. A. (1995). An empirical test of the cultural capital hypothesis. *Review of Black Political Economy, 23*(4), 1–27.

Junakar, P. N., Paul, S. & Yasmeen, W. (2004, June). *Are Asian migrants discriminated against in the labour market? A case study of Australia* (IZA Discussion Paper No. 1167). Bonn, Germany: Institute for the Study of Labor.

Kahn, L., & Shearer, D. (1988). Racial differences in professional basketball players' compensation. *Journal of Labor Economics, 6*(1), 40–61.

Kain, J. D. (1968). Housing segregation, Negro employment, and metropolitan decentralization. *Quarterly Journal of Economics, 82,*175–197.

Keith, V., & Herring, C. (1991). Skin tone and stratification in the black community. *American Journal of Sociology, 97,* 760–778.

Krashinsky, H. A. (2004). Do marital status and computer usage really change the wage structure? *Journal of Human Resources, 29*(3), 774–791.

Lang, K. (1986). A language theory of discrimination. *Quarterly Journal of Economics, 101*(2), 363–382.

Leonard, J. (1984a). Anti-discrimination or reverse discrimination? The impact of changing demographics, Title VII, and affirmative action on productivity. *Journal of Human Resources,19*(2), 145–174.

Leonard, J. (1984b). Impact of affirmative action on employment. *Journal of Labor Economics, 2*(4), 439–463.

Levine, R., Levkov, A., & Rubinstein, Y. (2008). *Racial discrimination and competition* (NBER Working Paper No. 14273). Cambridge, MA: National Bureau of Economic Research.

Loury, G. C. (1998). Discrimination in the post–civil rights era: Beyond market interactions. *Journal of Economic Perspectives, 12*(2), 117–126.

Lundberg, S., & Startz, R. (1983). Private discrimination and social intervention in competitive labor markets. *American Economic Review, 73*(3), 340–347.

Margo, R. A. (1992). Civilian occupations of ex-slaves in the Union army, 1862–1865. In R. W. Fogel & S. L. Engerman (Eds.), *Without consent or contract: Markets and production: Technical papers, Vol. 1* (pp. 170–185). New York: W. W. Norton.

Munnell, A. H., Tootell, G. M., Browne, L. E., & McEneaney, J. (1996). Mortgage lending in Boston. *American Economic Review, 86*(1), 25–53.

Nardinelli, C., & Simon, C. (1990). Customer racial discrimination in the market for memorabilia: The case of baseball. *Quarterly Journal of Economics, 110,* 575–595.

Peoples, J., & Robinson, R. (1996). Market structure and racial and gender discrimination: Evidence from the telecommunications industry. *American Journal of Economics and Sociology, 55,* 309–326.

Peoples, J., & Saunders, L. (1993). Trucking deregulation and the black/white wage gap. *Industrial and Labor Relations Review, 47,* 23–35.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review, 62,* 659–661.

Ransford, E. (1970). Skin color, life chances, and anti-white attitudes. *Social Problems, 18,* 164–178.

Robb, A., & Fairlie, R. (2007, January). *Determinants of business success: An examination of Asian-owned businesses in the United States* (IZA Discussion Paper No. 2566). Bonn, Germany: Institute for the Study of Labor.

Rose, N. L. (1987). Labor rent sharing and regulation: Evidence from the trucking industry. *Journal of Political Economy, 95,* 1146–1178.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C. A., Goeken, R., Hall, P. K., et al. (2008). *Integrated public use microdata series: Version 4.0* [Machine-readable database]. Minneapolis: Minnesota Population Center. Retrieved October 20, 2008, from http://usa.ipums.org/usa

Stewart, M. B. (1983). Racial discrimination and occupational attainment in Britain. *Economic Journal, 93*(371), 521–541.

Thernstrom, S., & Thernstrom, A. (1997). *America in black and white: One nation, indivisible.* New York: Simon & Schuster.

Turner, C. S., Tamura, R., & Mulholland, S. E. (2008). *Productivity differences: The importance of intra-state black-white schooling differences across the United States, 1840–2000* (MPRA Paper No. 7718). Munich: University Library of Munich, Germany.

U.S. Census Bureau. (2000). *Census 2000.* Available at http://www.census.gov/main/www/cen2000.html

U.S. Census Bureau. (2008a). *Current population survey, annual social and economic supplement.* Available at http://www.census.gov/cps

U.S. Census Bureau. (2008b). *Historical income tables: People* (table P-36). Available at http://www.census.gov/hhes/www/income/histinc/incpertoc.html

Vars, F. E., & Bowen, W. G. (1998). Scholastic aptitude, test scores, race, and academic performance in selective colleges and universities. In C. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 457–479). Washington, DC: Brookings Institution.

Welch, F. (1976). Employment quotas for minorities. *Journal of Political Economy, 84*(4, Pt. 2), S105–S139.

Wilson, W. J. (1996). *When work disappears: The world of the new urban poor.* New York: Knopf.

Zeng, Z., & Xie, Y. (2004). Asian-Americans' earnings disadvantage reexamined: The role of place of education. *American Journal of Sociology, 109*(5), 1075–1108.

# 56

# Economic Analysis of the Family

Steven Horwitz
*St. Lawrence University*

One of the notable characteristics of economics over the end of the twentieth and start of the twenty-first century is the extension of economic analysis to subject matter that was not traditionally thought of as economic. Books like *Freakonomics* (Levitt & Dubner, 2005) demonstrate these new applications of economics, and these books' popular success indicates that there is a demand for the use of economics to shed light on a variety of social issues. The origins of this extension of economics to the noneconomic in this fashion are often associated with work at the University of Chicago, in particular the work of 1992 Nobel Laureate Gary Becker. Much of Becker's (1960, 1973, 1974) seminal work in economics was devoted to showing how "the economic way of thinking" could enhance our understanding of a wide variety of social phenomena. One of the first social institutions Becker explored in this fashion was the human family.

In the almost 50 years since that first contribution, the economics of the family has exploded as an area of economic research. A whole variety of family-related phenomena, from how people choose marriage partners to how families make decisions about market versus household production to how household production is divided to how many children couples have to why and how frequently they get divorced to large-scale issues about the evolution of the form and function of the family, have all been subject to economic analysis. This literature has grown substantially in the last few decades, and this entry should be seen as an overture to the much richer work cited in the references.

## Theory

The economics of the family has explored almost every aspect of what might be termed the *life cycle* of the family: the time from marriage to divorce or death. In each case, the general strategy is similar in that it makes use of the economic way of thinking to analyze the choices that men and women make about the formation, continuation, and dissolution of families. At its most basic, the economic way of thinking proceeds as follows:

1 Identify the relevant decision makers and attempt to assess the costs and benefits they face in making the decisions in question, recognizing that both costs and benefits may have a large subjective component to them.
2. Note that the relevant costs and benefits are on the margin and that costs may be opportunity costs rather than explicit ones.
3. See whether the empirical data are consistent with predictions that emerge from the specification of the marginal costs and marginal benefits.

With respect to the family, this means that the economist is trying to understand decisions from marriage to childbirth to divorce as being the outcome of marginal benefit versus marginal cost comparisons by the people involved. When economists first started analyzing family issues this way, many objected that it was an attempt to reduce the romance and mystery of marriage and family down to cold calculations, especially financial ones. A further look at how the economics of the family proceeds shows that these concerns are largely misplaced.

The economics of the family simply argues that like all other human decisions, considerations of costs and benefits matter for decisions about the family, and the relevant costs and benefits need not be construed as narrowly financial or pecuniary. It is the focus on the margin and the comparison of costs and benefits that makes it an economic analysis, not any specification of what those costs and benefits might be.

The simple example of choosing to get married illustrates many of these points. First, how do spouses decide that each is the one? Conventionally, one might tell stories of love at first sight, or stars in their eyes, or a sign of some sort. All of those might be part of the decision, but at the very least, the lovers must consider whether someone "better" is out there. After all, rarely do we find perfection in a partner. Why stop looking now and choose this person to marry? From an economic standpoint, we might raise a number of considerations.

The love that spouses have for one another is a significant benefit from deciding to get married. They do get to spend the rest of their lives with one another. There are more material benefits to marriage, or more precisely to household formation. For the purposes of this chapter, we will assume that marriage and household formation happen together, even though in reality the latter sometimes comes first, which is in itself a question in the economics of the family: What are the costs and benefits of living together versus marriage? If the couple marries and moves in together, the spouses benefit from economies of scale (lower average costs of production). For example, if each one has a vacuum cleaner and a washer and dryer, they no longer need two of each. It is also cheaper, per person, to cook for two people, not to mention saving one set of utility payments.

On the cost side, prospective spouses have to consider the down sides of sharing space and resources. One of those costs is *imperfect preference satisfaction.* When decisions are made jointly, one or both partners frequently do not get exactly what they would like, especially as compared to each making the decision alone. The joint decision about what car to buy is often a good example here, because neither party gets exactly what he or she wants, as the number of people driving minivans they would not have bought on their own suggests. Compromise is a cost of marriage. Marriage is also a long-term legal commitment that is not cheap to end. Finally, there is the opportunity cost question: Could one have done better? Analysts of the economics of marriage have constructed search cost models that attempt to show at what point it no longer makes sense to keep looking for someone better, much like the way in which drivers at some point decide they do not think there will be a cheaper gas station up the street and just take the one at the next corner.

Economics can also be used to analyze changes in the frequency of marriage overall. For example, economists know that marriage rates have fallen over the twentieth century, and they know that people have been getting married later and later in life over the last 50 years. One way of looking at both phenomena is to ask whether the net benefits of marriage have fallen, especially for younger people, in recent decades. Historically, when women's economic opportunities were fewer and less well paying, the benefits to them from marriage were greater. Being able to share the much higher income of a husband was the key to economic survival for many women, certainly so if they wished to be able to support any children they might have.

Even for men, the benefits of marriage have fallen in modern times. To see why, one needs to examine the concept of household production. One way of conceiving the household that marriage creates is that it has a series of outputs it produces, and the members of that household are the human capital of the production process that leads to that output. The outputs of household production include everything from children and child rearing to cooking and cleaning to managing the household finances. Assuming the household is not completely self-sufficient, it will also require resources from outside the household (what is often termed *market production*) to be combined with household labor to produce those outputs. As with any other production process, household production involves a division of labor, in this case between the spouses.

Over time, the costs and benefits of marriage have changed. In particular, as women have become more equal participants in the labor market and as less labor has been required for household production, thanks to increased technology and more widely available and reasonably priced market substitutes, the benefits related to the household division of labor have fallen substantially. Women no longer need access to men's market incomes, and men no longer need someone to manage the household. Because of microwaves, automatic clothes washers and dryers, the much lower cost of dining out or using a dry cleaners, and the widespread use of child care providers, households produce fewer goods and services themselves and require less labor in the process. These changes in the costs and benefits of marriage have in turn led to different choices by men and women. These changes can help explain why one sees later and fewer marriages.

One of the important contributions to the economics of marriage is the idea of *assortative mating.* If one assumes that individuals are net benefit maximizers and that the marriage market has enough participants (both male and female), theory predicts that we will get assortative mating, which means that people will tend to marry those who bring similar levels of benefits to marriage as they do. Economists use the terms *high* and *low benefits* to refer to the human capital of the potential marital partners. High-benefits partners are those who bring high earnings potential, higher education (implying, perhaps, more interesting and desirable consumption preferences), and good health. In general, a high-benefit partner will do better marrying another high-benefit partner than a low-benefit one, unless there is a very unequal distribution of the total benefits of the marriage. Given the competitiveness of the marriage

market, that outcome is unlikely. And because high-benefit people are so likely to marry each other, there are few high-benefit people left for low-benefit people to marry; hence, they tend to marry other low-benefit people, resulting in the pattern predicted by assortative mating.

Economic theory can also say something about the decision to have children and how many. Prior to economics examining this question, there were few rigorous examinations of how fertility decisions got made. Once again, the logic of costs and benefits and the margin are central to the analysis. Children clearly provide their parents with benefits, both psychological (or emotional) and economic. Having a child can be a great source of joy to parents. Children also have value as economic assets, particularly as a source of support for parents in their old age. Children can also provide labor for market or household production. All of these benefits must be weighed against the costs. Obviously, children must be fed, clothed, and educated, each of which requires explicit monetary costs from the parents. Children also demand much of the parents' time, reducing the time they have for market production, other forms of household production, or leisure. Having children is a loss of freedom for the parents, because their choices now must take into account the effects on the child. Finally, there are economies of scale in child production. The average cost of raising a child declines with each successive child, because the marginal costs of additional children are declining. It clearly does not require twice as much labor to raise two children as it does one, and things like clothes and toys can often be passed down to the younger child.

Even choices about raising children can be understood using economics. If parents wish to discourage problematic behavior in children, thinking in terms of incentives, costs, and benefits can be very effective. Take the case of children who forget to bring their lunches to school. The parents can simply bring the children's lunches whenever this happens. Soon, the children will recognize that they can impose the costs of their own mistakes on the parents with the parents' cooperation, which dramatically reduces children's incentive to remember to bring their lunches. The parents, in this case, would like to find a way not to bear the costs of the children's decision. Specifically, one might wish to strive for situations where the parents are indifferent to the children's decision to remember or forget their lunches. The simplest way to do that is to refuse to bring the lunches. If the children remember them, great. If the children forget, the parents bear no cost; the children will survive a day without lunch and will learn that remembering is their responsibility and that they will pay the costs of not remembering. In essence, the parents have used responsibility as a proxy for the role played by property rights in economic analyses. By delineating what is whose, property rights set up spheres of responsibility, and parental assignments of responsibility do as well. As long as parents are willing to stick to those assignments, they can prevent children from imposing costs on them and creating the inefficiencies that go with those costs (Wittman, 2005).

One of the most famous economic explanations of parent–child relationships is Gary Becker's (1991) Rotten Kid Theorem. He argues that if parents are altruistic and wish to help their children, those children will act in ways that maximize family income. Put differently, we might expect that a child whose parents want to help him or her would choose to shirk obligations to the whole family and attempt to live off the parents' generosity. Becker's analysis suggests that this expectation is wrong. The child's degree of selfishness turns out not to matter, because even the most self-interested child will still take into account the effect of his or her actions on the rest of the family, effectively internalizing all possible externalities. As Becker points out, this theorem can be applied beyond the parent–child relationship to a variety of interactions within the family. He also notes that this does not mean that families will be without conflict. All the theorem predicts is that all have an interest in maximizing total family income. How that income is distributed among beneficiaries can still be the source of much conflict.

The economics of divorce is in many ways the mirror image of the economics of marriage. When the benefits to marriage fall, ceteris paribus, the opportunity cost of divorce falls and one sees more divorces; when the benefits to marriage rise, one should see fewer divorces. If couples gain less from the specialization that the household division of labor involves, then the incentive to stay married in the face of any sort of dissatisfaction with the marriage—especially of a psychological or emotional nature—is that much less. The legal environment matters a great deal here as well, because there are significant transaction costs to a divorce (much more so than a marriage). If the law makes divorce costly and difficult, one should see fewer divorces, even if couples are unhappy. To the extent the law reduces the transaction costs associated with divorce, divorce rates will rise, all else equal. As some have noted, these legal rules are not completely external to the economics of marriage. If the economic benefits to marriage were falling, one would expect that the margin of unhappiness that would produce a possible divorce would be lower as well (i.e., people will put up with less unhappiness if they are not getting other benefits from the marriage). In turn, this would produce a greater demand for divorces, which might well pressure political and legal institutions to reduce the transaction costs of divorce. A complete economics of the family has to account for the possible endogeneity of the institutional framework within which individuals decide.

## Applications and Empirical Evidence

Understanding some of the basic theory behind the economics of the family can help economists make sense of a great deal of history and current economic and demographic data. The evolution of the Western family over the last several hundred years illustrates a number of

the core economic principles discussed in the prior section. The transition from agriculture to industry is particularly instructive.

Several features of the family in the era of agriculture are worth noting. First, marriage was predominantly based on economic considerations rather than emotional ones. Marriage was necessary for survival, so finding a good mate was much more akin to finding a good work partner than anything else. Love, as we understand it today, was much more the province of the very, very few who did not have to worry about day-to-day survival. Second, the household was run by the husband. In this world, there was no real distinction between market and household production, because even crops that were sold on the market were produced by the household. Although women still disproportionately labored at what one today would call household production, they, along with children, were also expected to help with the crops or the cattle. And all of this was under the direction and supervision of the man. Third, children were viewed as economic assets. They were needed to contribute to production. This explains why families in the past (and families in agricultural areas today) tended to be larger. With falling marginal costs of children and potentially rising marginal benefits for the first several children, having a family of five, six, or seven made much more economic sense.

Industrialization changed the costs and benefits facing a number of family-related decisions and thereby changed the way families look and function. The crucial change was that the development of factories meant the physical separation of market and household production as men went out to work to earn an income, eventually leaving women at home in charge there. This heightened the division of labor within the family, which reached its peak in the late-nineteenth-century concept of men's and women's separate spheres. By ending the economic partnership component of marriage, industrialization was also a catalyst for the development and spread of love-based marriages. Finally, industrialization led to higher wages, which eventually allowed children and women to get out of the factories and into the homes, once their incomes were not necessary for survival and comfort. One result of this development was that children progressively lost their roles as economic assets and acquired new costs, because they could now devote time to education rather than production. This shift in the relative costs of children led to the (still ongoing) reduction in the size of the family and the related development of increasingly reliable birth control technology. Many of the features we associate with the modern Western family are products of the transition from agriculture to industrialization and the way in which it altered the costs and benefits facing parents. Economic theory is extremely useful in elucidating the process by which that evolution occurred.

As noted briefly earlier, one of the notable demographic trends of the last 40 or 50 years has been the decline in the marriage rate and the rise in the median age of first marriage. Between 1950 and 2000, the percentage of women aged 20 to 24 who were never married jumped from 32.3% to 72.8%, with the male numbers being 59.0% and 83.7%. For those aged 25 to 29, one sees a tripling of the percentage of women never married (38.9% in 2000) and a more than doubling of the male percentage (51.7% in 2000). The rates for those aged 30 to 34 also more than doubled in that 50-year period. The median age of first marriage for men rose from 23.0 to 26.9 and for women from 20.3 to 25.1 over roughly the same 50-year period. The overall marriage rate fell from 10.7% to 8.2%, and consistent with what theory would predict, birthrates over that period fell by 50% (Jacobsen, 2007).

Economic theory offers several possible explanations for the changes in the marriage rate, all of which emerge from the basic insight that marginal benefits and costs matter when people make marriage decisions. Women's greater productivity in the market has reduced the differences in comparative advantages between men and women, thereby reducing the benefits from the elements of specialization and exchange that characterize marriage. As men's and women's human capital look increasingly similar, the benefits of marriage fall while the costs remain roughly the same. This tendency may be somewhat offset by elements of assortative mating in that as men's and women's human capital converge, their tastes and the opportunity cost of their time converges as well, which might lead to more complementary consumption preferences, which would in turn increase the benefits of marriage. As marriage has evolved from narrowly economic to more about emotional and psychological factors, concerns about complementarities in production have gradually been replaced by concerns about complementarities in consumption. Signals about what one reads, watches, listens to, or eats are becoming far more relevant to the marital decision than what one does for a living or one's preferences about household or market production.

Men's incomes relative to women have fallen, and this is another potential factor. Men may delay marriage until they have levels of income that are, in their minds, sufficient in comparison to that of potential mates, perhaps because they wish to ensure a certain level of power in the household. A smaller income gap between men and women would mean it might take more time for men to get to that threshold, thereby leading to later first dates of marriage by men.

Two longer run social changes are of relevance here as well. Aside from the decline in the direct benefits of marriage, family formation may offer fewer benefits than in the past because more and more of the social functions that were once met by the family are now met outside of it. Other social institutions have, over time, taken on more of the economic, educational, religious, and social functions of the family, leaving the family with the core developmental and socialization functions (though market substitutes such as paid child care exist here too). These changes mean that families are less important to achieving important social goals, reducing the benefits of marriage. At the

same time, substitutes for marriage are less costly than they used to be since social disapproval of nonmarital cohabitation has almost disappeared, allowing couples (including same-sex ones) to get many of the remaining benefits of household formation and family outside legal marriage.

Countertrends for each of these explanations have also been identified in the literature, because respecification of the relevant costs and benefits can produce different predictions. For example, the reduction in the social functions of the family might make marriage more attractive to women who wish to pursue a career, do not want to engage in much household production, and would prefer having a committed companion to share valuable leisure time. As with other areas in economics, the economics of the family continues to explore which costs and benefits seem to be the most powerful in explaining observed outcomes.

The decline in birthrates also reflects powerful economic factors. The shift from agriculture to industry and the corresponding shift of children from being net producers to being net consumers are clearly borne out in the data. At the same time, the development of more reliable contraception, most likely an endogenous response to children becoming increasingly a net cost to parents, and changes in women's expectations about their own participation in market production have contributed to the fall in birthrates. In addition, reductions in infant and child mortality have made it possible for parents to have a given number of children survive to adulthood with fewer pregnancies and births. The decline in the birthrate may reflect not just a desire for fewer children but an increase in the efficiency of the cycle from pregnancy to adulthood. In addition, children were for many years a source of support for adults in their old age. With higher levels of wealth, parents are more able to save for their own old age, particularly through organized retirement plans, both public and private. This also reduces the benefits of having children.

Models of child production decisions often treat the decision as one about child-based consumption. That is, parents wish to produce a certain combination of the number of children they have and the investment in each of those children that delivers a multiplied total of child-based consumption. The investment can take the form of either parental time or market-acquired goods and services. In these models, the growth in parental wages over the last several decades has both income and substitution effects on child-based consumption. Rising income will lead parents to want more of it. However, rising wages imply rising opportunity costs of having children, which will reduce the amount of child-based consumption being sought because parents substitute less-time-intensive forms of consumption than those involving children. However, the increase in wages can also lead to families substituting market goods and services for their own time in the production of child-based consumption (e.g., hiring a nanny, providing children more toys to keep them entertained after school, or even sending older children to a boarding school). It might also be cheaper as wages rise to increase parents' child-based consumption by investing more resources in a small number of children rather than having more children, even given the economies of scale in child production.

The empirical evidence on the size of some of these effects varies. Empirical studies in the United States and Europe indicate that a 10% increase in women's wages will produce anywhere from an 8% to a 17% decline in births, depending on the wealth of the region being studied, while a 10% increase in men's wages would increase births by anywhere from 10% to 13% (Winegarden, 1984). The negative elasticity associated with women's wages reflects the substitution effects of women's time, while the positive elasticity of men's wages reflects the pure income effect that a higher male wage has for his wife. Even with correlation data such as these, causality remains controversial: Do women have fewer children because they are working more, or are they working more because they are having fewer children? Economists do not have a very good understanding of how long-run changes in the economy affect social choices, and standard econometric studies can show only correlation, not causation.

An additional application of the basic economic theory of the family regards the division of labor within the household. The standard model argues that the spouse with the lowest opportunity cost of his or her time as measured by their market wage should specialize in household production, while the other should specialize in market production. The model assumed that differences in wages between men and women were large enough that they mattered for such purposes. The model would also predict that as male–female wage differentials disappeared, one should see a more even distribution of work in the household. As women's wages rise, one should see men taking on a more equal share of household production.

Empirically, we have seen convergence in the time men and women spend on household production, but not to the same degree as the convergence in wages. That is, women still do a disproportionate (to the opportunity cost of their time) amount of the household production. Recent empirical work shows that men and women with roughly the same human capital and the same demographic features (e.g., age and marital status) working in the same job will earn close to the same wage. However, married men and women still do significantly different, though less so than in the past, amounts of housework. In 1965, the average for married men with a child aged 5 or over was 5.3 hours of housework per week, while women averaged 30.3 hours. By 2004, the men in that group averaged 9.5 hours per week, while the women fell to 16.9 hours. In relative terms, women went from doing almost six times the housework to almost twice as much. There is no question that the economic model can explain a good deal of this, but the disparity by gender remains, with women putting in 7.4 more hours per week, despite the fact that the wage differential is nowhere the 2:1 ratio.

It is also worth noting that the total amount of time spent on housework fell from 35.6 hours per week to 26.4 hours (all data as reported in Jacobsen, 2007, p. 111). That overall reduction is due to a combination of better household technology and cheaper market substitutes, as well as lower standards of cleanliness.

Even if one takes away the children at home, the men do only 0.7 hours more per week, and the women do 1 hour less, leaving a differential of 5.7 hours. The explanation for the difference cannot be the presence of children alone. A complete explanation for the remaining difference will likely focus on factors giving the male more bargaining power in the discussion over the division of housework, perhaps deriving from the higher cost of exiting the relationship for women, particularly when there are children involved. Economists have used a variety of bargaining models to explore both marital choice and the division of household labor (e.g., Lundberg & Pollack, 1993).

The economics of divorce provides yet another application of the basic model. Divorce rates in the United States are notably higher today than 40 years ago, despite a slow downward trend since about 1980. After a long-run slow increase since the early twentieth century, interrupted by a spike at the end of World War II and a leveling off for the 20 years afterward, the divorce rate more than doubled in the period between the mid-1960s and the early 1980s. Two of the most straightforward explanations of the long-run increase are the increased ability of women to survive economically outside of a marriage, discussed earlier, and the reduction in the birthrate. Both of these reduce the marginal cost of divorce. The first does so by reducing the financial burden on women and the second by reducing the number of marriages with any children, or with minor children, at a given time. Married couples with minor children are more likely to try to stick it out in a bad marriage, so anything that reduces the number of such children lowers the cost of divorce, inducing more divorce.

The large jump from the mid-1960s until the early 1980s and the relatively high plateau maintained since then require additional explanations. The mid-1960s date is not accidental, because it reflects the beginning of the loosening of divorce law across the country. In particular, the advent of no-fault divorce is often credited as a key factor in the rising divorce rate of that period. This change in divorce law allowed couples to divorce without having to prove adultery, abuse, or abandonment by the spouse. They could simply decide they did not wish to be married anymore. In most states, by the late 1970s, the unilateral desire by one partner to leave the marriage was sufficient to get a divorce. The controversy in the literature is the degree to which no-fault was a cause or consequence of a rising desire for divorce. Certainly, by lowering the cost of divorce, no-fault made divorce more likely, but others argue that it was a shift in the demand for divorce that led to the changes in the law. In fact, the divorce rate does seem to start its climb before no-fault had really spread to most states, lending credence to the idea that it was the intensification of the longer-run factors noted previously that pushed states to liberalize their divorce laws. Once liberalized, however, the lowered cost induced a movement along the demand-for-divorce curve, causing another jump.

A second explanation for that increase and sustained plateau is that there are expectation-matching and signaling problems as men and women adjust to new norms (Allen, 1998). Couples who married under the mutual assumption of a more strict division of labor by gender may well have found themselves 15 years later in a world where the wife was working, leading to tensions in the marriage because the reality did not match their expectations. The jump in divorces that quickly followed the advent of no-fault suggests that there were latent divorces waiting to happen, and these might have been one type. The signaling problems reflect the uncertainty that has come with changing social norms and gender roles. By the 1970s, couples were less likely to expect a strict division of labor, but men in particular may have had a difficult time in determining whether a prospective spouse was likely to want a career or to stay home. Rather than a flat-out wrong expectation, the signals about what expectation to have became increasingly noisier, making it more likely that mistakes would be made. Even for women, determining whether a potential husband would be accepting of their deciding later on to take up a career became more difficult in the new environment. This would likely generate more bad matches that would reveal themselves down the road, which is consistent with the ongoing high divorce rate.

This signaling theory should also predict a decline in the divorce rate, at least on the margin, as social norms become more predictable. Recent divorce data suggest this might be happening. The divorce rate has been very slowly falling, and when it is looked at from a cohort perspective, we find that divorce rates among those married just a few years have fallen more notably. A decade ago, couples married just 5 to 10 years were more likely to get divorced than couples today married that long. An explanation for this is that couples are making better matches today because they are facing less-noisy signals in the marriage market. Men and women have more similar life goals and are more tolerant of the need to bend their own goals to those of spouses. Even as the marriage rate continues to fall, the most recent evidence suggests that those marriages that are happening are more likely to stay together than in years past (U.S. Census Bureau, 2007).

## Policy Implications

The economics of the family can inform an analysis of various policy proposals. In this section, three such issues are examined using the approaches outlined in previous sections.

One of the more fascinating questions worth exploring is how policy choices can affect women's labor force participation decisions. As an example, consider the way in

which the U.S. tax system treats secondary earners. The typical economic model of the family assumes that labor force decisions are made sequentially—that is, the higher-earning spouse is assumed to work in the market, and given that person's decision, the family decides whether the lower-earning spouse should also work in the market and how much. More often than not, the secondary earner is the woman. Tax policy enters into the picture because married couples are almost always better off filing jointly than using the IRS's *married, filing separately* category, even given the discussion to follow. Filing jointly means that the secondary earner's first dollar of income is taxed at the marginal rate applicable to the primary earner's last dollar. So rather than paying lower rates on portions of the secondary earner's income, the couple filing jointly sees a much larger portion of all secondary income taxed. When that effect is combined with the law's refusal to treat child-care payments as an expense of employment, the incentives for the secondary earner in a family with small children are such that the disposable income from the secondary earner's work is actually quite low. Some analysts (McCaffery, 1999) have suggested that if married couples could file truly separately (as distinct from married, filing separately, under current law), with each one's income being taxed separately, this problem would be substantially lessened.

Analyzing the policy of no-fault divorce is another area in which the economics of the family can contribute. Historically, a spouse had to show abuse, abandonment, or adultery in order to get a divorce. For couples who were just unhappy, this often meant having to lie to have the grounds necessary for divorce, and it placed one spouse in the position of being at fault for the divorce. From an economic perspective, this process involved rather high transaction costs for unhappy couples. In the late 1960s and early 1970s, more states moved to so-called no-fault divorce, by which one or the other spouse could terminate the marriage for whatever reason he or she desired. This policy is probably more accurately called *unilateral* divorce, because the divorce does not require the consent of both parties. One advantage of so-called fault-based divorce is that couples who were unhappy at the very least had to cooperate in a lie to get a divorce. The need to cooperate in a lie meant that the divorce could take place only if both parties were willing to take affirmative steps to give up on the marriage. This may well have made for fewer adversarial divorces and greater ease in dividing up assets because both parties thought they would be better off separated. Today, the decision is unilateral in most states and therefore does not require any cooperation of this sort.

The problem with no-fault is that unilateral divorce puts the weaker party, usually the woman, at the mercy of the stronger. Imagine the wife who has given up her own career to help her husband's, to raise children, or both. If he decides he wants a divorce, she has very little leverage to prevent him or to negotiate reasonable terms (Cohen, 1987). Prior to no-fault, she could at least hold out on cooperating to convince a judge that there were sufficient grounds. Or if one imagines a divorce regime where mutual consent replaces unilateral desire or abuse, abandonment, or adultery as the grounds for divorce, the financially weaker party gains significant bargaining power.

This can perhaps best be seen by using the Coase Theorem, one implication of which is that if transaction costs are sufficiently low, any situation involving external costs can be renegotiated to make the harmed party whole. What a mutual consent divorce regime would do is force couples into a negotiation process that would ensure that the weaker party is able to make a credible claim to compensation for the breaking of the marital contract. This effect of such a policy is a clear benefit. However, this shift would come with costs as well. One of the advantages of unilateral divorce is that it makes it very easy for a harmed party to exit the marriage. Imagine a situation where a wife was deeply unhappy but not abused, nor was there any adultery. Under unilateral no-fault, she could file for divorce and get out. If she were in fact being abused, she would not need to demonstrate to a judge that abuse was present. By raising the costs of divorce, going to a regime of mutual consent does make it harder for the victimized spouse to lose out economically, but it also makes it harder to get out of a bad marriage in the first place. An economic analysis of the family can help policy makers think through the trade-off here and try to find a set of policies that would make it easier for victimized spouses to exit bad marriages but to do so in ways that do not leave them substantially worse off.

The economic approach to child production also sheds light on policies designed to reduce birthrates in parts of the world where the population is thought to be growing too quickly. In the past, policy makers often focused on improving education and access to birth control—for example, by distributing condoms in third world countries. The assumptions were that population was, in fact, too high and that parents wanted smaller families but simply did not know how to make that happen or lacked the resources to do so. The economic approach suggests that all of these assumptions may be faulty. It may well be the case that parents in the third world are having the number of children they actually desire to have, given the circumstances in which they find themselves. The marginal benefits of additional children often do outweigh the costs in poorer societies for reasons this chapter has already touched on. At the very least, the economic approach to child production leads analysts to ask a different set of questions about this situation. Many have noted that freely distributed condoms often do not get used for their intended purpose and end up being used for other things. The economics of the family can explain that by asking whether the reason for large families in these societies is not ignorance or lack of resources but a reasonably rational calculation of the benefits and costs of child production. That population growth rates in the developed world are different might just reflect a difference in

those costs and benefits, rather than education or access to resources per se.

These are but a few examples of how economic analyses of the family can inform policy discussions. Too often, family decision making is understood as standing outside the province of rationality and the kind of marginal benefit and marginal cost thinking of economics. Ignoring the economic approach to the family can lead to bad policy decisions if families are, in fact, making decisions along the lines that economics suggests.

## Future Directions

The social institution of the family has changed enormously over the last 50 years, which is roughly the same length of time it has been a serious object of study for economists. There is little doubt that the change will continue in the years to come. Economists will likely be kept very busy trying to understand the continued evolution of gender relationships and the household division of labor, particularly as communications technology appears to be making it increasingly possible to work from home. Analyzing child production and raising decisions in a world where genetic choice might be possible and where the cost of raising kids continues to climb will provide another set of challenges. The same can be said of the developing world as globalization brings economic pressures on family forms and gender roles there. Finally, as same-sex marriage remains on the policy agenda, and as more same-sex couples create households and raise children, economists may find interesting research questions there because such households face the same decisions as opposite-sex ones have over the years.

## Conclusion

Exploring the social institution of the family through the eyes of economics not only demonstrates the variety of phenomena that can be examined using the economic way of thinking, it also provides a different and valuable perspective on how families function and the choices that individuals make within them. Economics brings into relief the constraints under which family members choose and the various marginal benefits and marginal costs that might inform their decision making. Seeing the family in this light can help us make sense of some of the major demographic changes of the last few generations as well as inform policy makers as to the likely consequences, both intended and unintended, of various family policy proposals. Like all other social institutions, the family exists to reduce transaction costs and to solve human coordination problems. Economics helps us identify exactly how families do so and thereby informs our attempts to improve them.

## References and Further Readings


Allen, D. W. (1998). No-fault divorce in Canada: Its cause and effect. *Journal of Economic Behavior and Organizations, 37,* 129–149.

Becker, G. S. (1960). An economic analysis of fertility. In Universities-National Bureau (Ed.), *Demographic and economic change in developed countries* (pp. 209–231). Princeton, NJ: Princeton University Press.

Becker, G. S. (1973). A theory of marriage: Part 1. *Journal of Political Economy, 81,* 813–846.

Becker, G. S. (1974). A theory of marriage: Part 2. *Journal of Political Economy, 82,* S11–S26.

Becker, G. S. (1991). *A treatise on the family* (Enl. ed.). Cambridge, MA: Harvard University Press.

Becker, G. S., & Lewis, H. G. (1973). On the interaction between the quantity and quality of children. *Journal of Political Economy, 81,* S279–S288.

Cohen, L. (1987). Marriage, divorce, and quasi rents: Or, "I gave him the best years of my life." *Journal of Legal Studies, 16,* 267–303.

Coontz, S. (2005). *Marriage, a history: From obedience to intimacy or how love conquered marriage.* New York: Viking.

Grossbard-Shechtman, S. (1993). *On the economics of marriage: A theory of marriage, labor and divorce.* Boulder, CO: Westview Press.

Hadfield, G. (1999). A coordination model of the sexual division of labor. *Journal of Economic Behavior and Organization, 40,* 125–153.

Jacobsen, J. P. (2007). *The economics of gender* (3rd ed.). Malden, MA: Blackwell.

Levitt, S. D., & Dubner, S. J. (2005). *Freakonomics.* New York: HarperCollins.

Lundberg, S., & Pollack, R. (1993). Separate spheres bargaining and the marriage market. *Journal of Political Economy, 101,* 988–1010.

McCaffery, E. J. (1999). *Taxing women.* Chicago: University of Chicago Press.

Parkman, A. (2000). *Good intentions gone awry.* Lanham, MD: Rowman & Littlefield.

Posner, R. A. (1992). *Sex and reason.* Cambridge, MA: Harvard University Press.

Robinson, J. P., & Godbey, G. (1997). *Time for life: The surprising ways Americans use their time.* University Park: Pennsylvania State University Press.

Rosenzweig, M. R., & Schultz, T. P. (1985). The demand for and supply of births: Fertility and its life cycle consequences. *American Economic Review, 75,* 992–1015.

Rosenzweig, M. R., & Stark, O. (Eds.). (1997). *Handbook of population and family economics.* Amsterdam: Elsevier.

Shorter, E. (1975). *The making of the modern family.* New York: Basic Books.

U.S. Census Bureau. (2007). *Survey of income and program participation (SIPP), 2004 panel, wave 2 topical module.* Retrieved February 2, 2009, from http://www.census.gov/population/socdemo/marital-hist/2004/Table2.2004.xls

Winegarden, C. R. (1984). Women's fertility, market work, and marital status: A test of the new household economics with international data. *Economica, 51,* 447–456.

Wittman, D. (2005). The internal organization of the family: Economic analysis and psychological advice. *Kyklos, 58,* 121–144.

# 57

# Economics of Aging

Agneta Kruse
*Lund University*

Population aging is a worldwide phenomenon. The demographic transition from a state with high fertility and mortality rates to a state with low fertility, low mortality, and increased longevity has changed the well-known population pyramid into what looks more like a rectangle or a skyscraper. This means fewer children and more old-aged people in relation to those of working age. Aging started in the industrialized world and has thus gone farther in these countries than in developing countries, where it started later. The transition is far from complete in the industrialized world, and aging is forecast to go on during the entire period of prognosis, up to the year 2050.

Population aging has caused anxieties about the economic effects of aging: the possibilities of supporting an increasing number of older persons with a shrinking workforce. This chapter focuses on the economics of aging in the industrialized world. The next section describes the demographic changes and aging, their history and prognoses of future development. The following section uses a life cycle approach to give a theoretical foundation for analyzing the economic aspects of aging, from both the individual's and society's points of view. Later sections analyze the effect of aging on production and consumption patterns. As old age pensions constitute such a large part of life cycle redistribution, special attention is given to pensions and the eventual strain caused by aging. The conclusions from these sections are that demographic changes will cause substantial economic pressure, ceteris paribus. The final section discusses possible remedies and concludes the chapter.

## Aging Populations

Historically, aging started with a drop in fertility rates, which had already started to decline in many industrialized countries in the nineteenth century. In 1950, the total fertility rate was 3.45 in the United States (meaning that on average each woman was expected to give birth to 3.45 children), 2.66 in Europe, and 2.75 in Japan; it had fallen to 2.04 in the United States, 1.41 in Germany, and 1.29 in Japan by 2000. A fertility rate below 2.1 is below reproduction rate; the population will eventually shrink if immigration does not compensate for the low fertility. There is a rather large dispersion in fertility rates between countries within Europe, with Italy and Germany having the very low fertility rates of 1.29 and 1.35, respectively, while Sweden has a fertility rate of 1.67 and the United Kingdom 1.70 (United Nations, 2007).

It takes some 20 years before a birth cohort enters the labor market. Thus, low fertility rates will give rise to successively smaller cohorts entering the labor force.

In the next phase, population aging was driven by a drop in mortality and increases in longevity. Between 1950 and 2000, life expectancy at birth increased by 8.5 years in the United States, 11.2 years in Germany, 13.9 years in Italy, and 18 years in Japan, and population prognoses forecast that life expectancy will increase by at least 5 years by 2050 (United Nations, 2007).

A way of measuring aging is by dependency ratios, where the child dependency ratio is the number of children in relation to the number of people of working age, and the old age dependency ratio is the number of people aged 65 or older in relation to those of working age. The development of these ratios is shown in Table 57.1.

**Table 57.1** Dependency Ratios

| | *Child Dependency Ratio*[a] | | | | | *Old Age Dependency Ratio*[b] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Region* | *1950* | *1975* | *2000* | *2025* | *2050* | *1950* | *1975* | *2000* | *2025* | *2050* |
| Europe | 40 | 37 | 26 | 23 | 25 | 13 | 18 | 22 | 32 | 48 |
| Sweden | 35 | 32 | 29 | 28 | 28 | 15 | 24 | 27 | 36 | 41 |
| Germany | 35 | 34 | 23 | 21 | 24 | 14 | 23 | 24 | 39 | 54 |
| Italy | 40 | 38 | 21 | 20 | 25 | 13 | 19 | 27 | 39 | 60 |
| United Kingdom | 33 | 37 | 29 | 27 | 27 | 16 | 22 | 24 | 32 | 40 |
| United States | 42 | 39 | 33 | 30 | 28 | 13 | 16 | 19 | 28 | 34 |
| Canada | 47 | 40 | 28 | 24 | 27 | 12 | 13 | 18 | 33 | 44 |
| South America | 69 | 71 | 49 | 34 | 28 | 6 | 8 | 9 | 16 | 29 |
| Australia and New Zealand | 42 | 44 | 32 | 28 | 27 | 13 | 14 | 18 | 31 | 41 |
| Africa | 76 | 86 | 78 | 61 | 43 | 6 | 6 | 6 | 7 | 11 |
| Asia | 61 | 71 | 48 | 34 | 28 | 7 | 7 | 9 | 15 | 27 |
| Japan | 59 | 36 | 21 | 19 | 22 | 8 | 12 | 25 | 50 | 74 |

SOURCE: United Nations (2007).

NOTES: Number of children or older persons per 100 working-age people from 1950 to 2000 and prognosis up to 2050, medium variant. (a) Ratio of population between 0 and 14 years of age to population between 15 and 64. (b) Ratio of population 65 and older to population between 15 and 64.

As can be seen in the table, aging started in most regions and countries with a drop in fertility; the child dependency ratio fell. In the next phase, population aging gave rise to increases in the old age dependency ratio. The table shows that aging has been going on for a long time and is forecasted to continue.

Table 57.1 shows the UN population forecasts with the medium variant. It should be emphasized that demographic forecasts are highly uncertain. For example, fertility rates have shown great variability in the past; the forecasts assume a constant fertility rate. Also, so far, the drop in mortality has constantly been underestimated in population forecasts, as have the increases in longevity. Even with great uncertainties in the forecasts, there is no doubt that the population is rapidly aging.

Aging brings forward changes in consumption and production patterns.

## Life Cycle Approach

No human being can support himself or herself by his or her own labor during a whole lifetime. Infancy is an obvious example, as is old age. Figure 57.1 sketches some important economic events in an individual's life: entrance into and exit from the labor market plus the length of retirement period. Individuals need to consume in periods when they do not have any income. To survive, individuals have to cooperate. The main institutions for cooperation are the family, the market, and the state, which exist side by side, and their importance differs in different countries as well as historically in the same country. The way of mixing the institutions and organizing them has effects on the pressure caused by aging, as is evident later on in this chapter; for a comprehensive analysis, see Lars Söderström (2008).

E is the time of entry into adulthood and the labor market. Childhood extends up to the point E. The period from E to EX represents working years; EX is the exit from the labor market into retirement. D is the time of death; the period between EX and D is the number of years as a retiree. During the working years, the individual earns an income, assumed to increase with age. Assume that the individual wants to maximize lifetime utility, subject to the budget constraint determined by lifetime income. The planning problem is then to transfer income from working years to nonworking years and to smooth consumption possibilities.

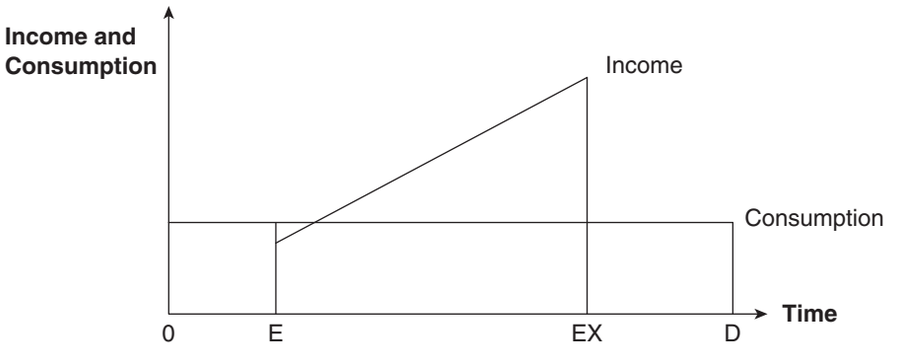


**Figure 57.1** Life Cycle Income and Consumption Planning Problem

The assumptions in Figure 57.1 are that income is increasing with age and that an even consumption stream over the years maximizes utility. Of course, other patterns are possible; they are discussed in the analysis of aging and the welfare state.

Figure 57.1 may also be used as an illustration of a society at a specific time. The number of children is then the number between origo and E, the number of people of working age is between E and EX, and the number of retirees is between EX and D. The income of those in working years is used for the support of all coliving individuals.

The figure highlights the planning problem faced by the individual and society—that is, how to organize transfers over years and between generations in an efficient and utility-maximizing way. The figure is drawn as if there were total certainty in life, the income stream is certain, the possibility of working until the age of EX is certain, and the time of death is certain, to mention but a few uncertainties that one meets in real life. Apart from childhood and old age, the risk of periods of work incapacity during working age, due to sickness or unemployment, for example, are examples of individuals' inability to always support themselves. The purpose of the institution for cooperation—be it the family, the market, or the state—is then not only to even out consumption possibilities over the life cycle, but also to function as an insurer against income losses, the risk of high consumption needs, like health care and old age care, as well as poverty.

The life cycle illustration gives an intuitive insight into the problems that an aging population might cause. The following sections analyze it further.

## Production and Labor Supply

So far we have shown a purely demographic view. However, more important from an economic point of view is the labor force in relation to the number of people being supported. At any given time, the number of people being supported is determined by the demarcation line between childhood and adulthood, between working ages and retirement, and finally between retirement and death. The UN figures used in Table 57.1 with the delimitation of age groups 0 to 14, 15 to 64, and 65 and older do not describe the economic-support reality in the industrialized world. In Western societies, there has been a trend of postponing entrance into the labor market. The expansion of higher education means that an increasing portion of each cohort studies at high school and university and delays entrance into the labor market. Besides, retirement age has shown a declining trend for several decades, even if it seems to have turned upward the in last decade. Economic growth is a common explanatory factor for both trends. Early retirement occurs both because pension systems have subsidized early retirement (see Gruber & Wise, 1999, for international comparisons) and because leisure is a normal good with positive income elasticity.

Furthermore, what is economically important is labor supply. Table 57.2 shows total labor force participation rates and the participation rates in different age groups.

**Table 57.2** Labor Force Participation Rates in Different Age Groups, 2007

| | *Ages 15 to 24* | *Ages 25 to 54* | *Ages 55 to 64* | *Ages 15 to 64* |
|---|---|---|---|---|
| Germany | 51.3 | 87.1 | 58.0 | 75.6 |
| Italy | 30.9 | 77.6 | 34.6 | 62.5 |
| Sweden | 57.1 | 90.0 | 73.0 | 80.6 |
| United Kingdom | 65.3 | 84.5 | 59.3 | 73.6 |
| European Union 15 | 48.8 | 84.6 | 48.8 | 72.5 |
| United States | 59.4 | 83.0 | 63.8 | 75.3 |
| Japan | 44.9 | 83.3 | 68.4 | 73.6 |

SOURCE: Organisation for Economic Co-operation and Development (OECD, 2008).

If the labor force is defined as people between 15 and 64 years of age, the participation rate is around three quarters in the countries shown in the table, far less in Italy, and somewhat higher in Sweden. Labor supply changes over the life cycle, with a rather low supply in young and old ages. The ages between 25 and 54 may be seen as the most productive years with high participation rates. The differences between age groups mean that the age structure and changes in it will influence a country's production capacity.

Using labor force participation figures for calculating what might be called an effective old age dependency ratio gives quite different and higher ratios than the ones shown in Table 57.1; see Table 57.3.

The use of the labor force instead of the number of persons increases the old age dependency ratio by 6 percentage points (for the United States) at the least and 18 percentage points (for Italy) at most—that is, quite substantial increases. Moreover, the use of population figures to forecast the old age dependency ratio in 2050 gives substantial increases in the ratio, as seen in Table 57.1. The effective old age dependency ratio exceeds these significantly; by the middle of this century, there will be one retiree per working person in Japan and almost one retiree per working person in Italy. However, even these figures are underestimates if we are interested in actual work performed; unemployed people are included in the labor force, as are people on sick leave, holiday, and parental leave. In addition, participation rates only partly capture the labor supply; they do not tell how many hours per week or how many weeks per year a person works, or how many years constitute a working life. All these aspects matter, and the chapter comes back to this when discussing pension systems.

These future prospects have given rise to numerous alarm reports, but these are often built on static calculations. It is worth emphasizing that economies are flexible and that there are adaptation possibilities. Prices will change, giving incentives to adapt to new circumstances. As an example, there is empirical evidence from cross-section and time-series data that a slower growth in labor force induces a more rapid productivity growth (Cutler, Poterba, Sheiner, & Summers, 1990) Furthermore, institutions matter; the way of organizing life cycle redistribution and support of older persons will most certain change.

The last 50 to 60 years have seen important changes in the labor market and in different groups' participation in the labor market. One trend is the increase in market work by women. There has also been a very marked downward trend in the participation rates by older men. Lately, this trend seems to be taking a turn upward. Gary Burtless (2008) shows that there is a trend of postponing retirement in industrialized countries; the age at retirement is increasing. Gabriella Sjögren Lindquist and Eskil Wadensjö (in press) discuss factors behind the decision to exit the labor market. They divide these into individual factors, such as health, family situation, and social and occupational pension schemes, influencing the supply side, and institutional factors, such as wage system and mandatory retirement age, influencing the demand for older workers. They conclude that the closing down of roads to early exit and the reformed pension system has delayed retirement in Sweden.

**Table 57.3** Old Age Dependency Ratios, 2005 and 2050

| | *2005* | | *2050* | |
|---|---|---|---|---|
| | *65 and Older / (Number of Persons)* | *65 and Older / (Labor Force)* | *65 and Older / (Number of Persons)* | *65 and Older / (Labor Force)*[a] |
| Germany | 0.28 | 0.37 | 0.54 | 0.71 |
| Italy | 0.30 | 0.48 | 0.60 | 0.97 |
| Sweden | 0.26 | 0.33 | 0.41 | 0.50 |
| United Kingdom | 0.24 | 0.33 | 0.40 | 0.55 |
| United States | 0.18 | 0.24 | 0.34 | 0.45 |
| Japan | 0.30 | 0.40 | 0.74 | 1.0 |

SOURCE: Author's calculations using figures from United Nations (2007) and OECD (2008).

NOTES: Age range for *number of persons* and *labor force* is 15 to 64. (a) Calculations are made with the assumption that labor force participation rates are the same in 2050 as those reported in Table 57.2.

### Aging and Capital

Aging may also influence savings and capital accumulation. Albert Ando and Franco Modigliani's (1963) hypothesis of consumption and saving over the life cycle assumes borrowing during childhood, positive saving during working ages, especially during the later phase of working ages, say between 50 years of age to 64, and dissaving during retirement. There is empirical evidence to support the theory—among others, Solveig Erlandsen and Ragnar Nymoen's (2008) empirical study of private consumption and capital accumulation using Norwegian data strongly supports the theory—even if some contrasting evidence also exists.

The population forecasts show that the number of people in working ages will decrease, which means that capital per worker will increase, other things being equal. If the capital–labor ratio was optimal at the outset, there will be too much capital; savings can be reduced and consumption increased. This is a conclusion drawn by, for example, Cutler et al. (1990). However, with fewer workers to support an increasing number of older persons, an increase in productivity caused by more capital per worker should be a welcome effect. The effect of aging on savings and capital is further analyzed in the discussion of pension reforms.

## Production and Consumption Patterns

Cutler et al. (1990) use what they call a support ratio to highlight changes in production capacity and changes in consumption patterns as the population ages. The support ratio is defined as

$$\alpha = LF / CON, \quad (1)$$

where LF is the effective labor force and CON is the effective number of consumers.

Cutler et al. (1990) start by assuming that LF and CON simply are defined as the number in different age groups—that is, an analysis very similar to the analysis discussed earlier called *pure* demography.

As pointed out in the introduction, aging started with a decline in fertility, which means that the support burden of children lessens. There is a time span before the old age dependency ratio increases; thus, there is a period in which the support burden is lessened. However, in the next phase of the demographic transition, the number of older persons increases, and so does the old age dependency ratio. The per capita consumption demands (needs) from older persons surpass by far that of children. A changing age structure, with fewer children and more older persons, does not have a neutral effect on the support burden. Cutler et al. (1990) therefore qualify the support ratio, taking different consumption patterns in different age groups into account as well as age-specific labor force participation rates. The support ratio will decline in the decades to come. One of their conclusions is that the support ratio is more sensitive to changing consumption patterns than to changes in the labor force participation rates (see also Erlandsen & Nymoen, 2008).

### Aging and the Welfare State

In many countries, a large part of life cycle redistribution takes the form of publicly provided or subsidized consumption and transfers. Division of public expenditure in accordance with the problems the welfare state aims to remedy gives essentially three categories: (1) life cycle redistribution, (2) risk insurance, and (3) what may be called redistribution in accord with our ethical preferences—that is, support of the genuinely weak. A comparison of expenditures in five countries, Britain, Hungary, Italy, Poland, and Sweden, in the middle of the 1990s shows that life cycle redistribution constituted by far the largest part. It ranged from 60% in Britain to more than 80% in Poland. The aid to the weak was the smallest part, below 10% in all the countries except in Britain, where it counted for just above one fifth (Kruse, 1997).

Of course, the importance of life cycle redistribution in the welfare state makes it sensitive to aging. During childhood, the individual is a net receiver of public expenditure; during working ages, a net contributor; and during old age, again a net receiver. During childhood, the main items are education, publicly provided or subsidized child care, and child allowances, to mention a few. During old age, the most important expenditure areas are public pensions, health care, and old age care. With a welfare state model, these expenditures are financed mostly by taxes, which give high tax wedges and are detrimental to employment.

There is widespread concern that population aging will put the welfare state under strain and force a retrenchment. With aging, the tax base decreases if the decline in the number of workers is not compensated for by productivity increases. At the same time, aging increases the demand for pensions, health care, and old age care. Prognoses of these important expenditure areas until the year 2050 are reported, for example, in EU Economic Policy Committee (EU, 2003). Table 57.4 shows these prognoses for a selection of European countries. Per capita expenditure is assumed to be constant; combining per capita expenditure with demographic changes, one gets the following expenditures.

Take health care as an example. The consumption of health care increases strongly with age. Björn Lindgren and Carl Hampus Lyttkens (in press) present figures showing the per capita cost for persons aged 80 and older to be around four times the cost for persons aged 30 to 50 years. This figure stems from a simulation model for Sweden. EU (2003) shows an even more marked increase with age. Note that taking this into consideration would change the consumption pattern in the stylized life cycle picture in

**Table 57.4** Public Expenditures on Public Pensions, Health Care, and Old Age Care as a Percentage of GDP in 2000 and Projections of Change From 2000 to 2050

| | *Pensions* | | *Health and Old Age Care* | |
|---|---|---|---|---|
| | *2000* | *Change From 2000 to 2050* | *2000* | *Change From 2000 to 2050* |
| Belgium | 10.0 | 3.3 | 6.1 | 2.1 |
| Denmark | 10.5 | 2.9 | 8.0 | 2.7 |
| Finland | 11.3 | 4.7 | 6.2 | 2.8 |
| France | 12.1 | 3.8 | 6.9 | 1.7 |
| Germany | 11.8 | 5.0 | 5.7[a] | 1.4 |
| Italy | 13.8 | 0.3 | 5.5 | 1.9 |
| Sweden | 9.0 | 1.7 | 8.8 | 3.0 |
| United Kingdom | 5.5 | −1.1 | 6.3 | 1.8 |

SOURCE: EU (2003).

NOTE: (a) Applies to health care only.

Figure 57.1. In all the countries in Table 57.4, the number of people in the age group 80 and older will increase rapidly. From 2000 to 2030, the 80 and older share in the population will increase from 4.0% to 9.0% in Italy, from 3.7% to 7.5% in France, and from 3.5% to 7.2% in Germany (United Nations, 2007).

To what extent the increased number of elderly people will actually lead to increases in the health care costs is, however, a matter of dispute. Friedrich Breyer and Stefan Felder (2006) report that attempts to foresee the increases in payroll tax rates to cover German health care costs give estimates with range as wide as between 16.5% and 39.5% in 2050. The lower figure stems from simulations using demographic change alone, while the higher one incorporates progress in medical technology, which has shown to be highly cost increasing. Although many technological advances are efficiency enhancing and cost reducing—using medicine instead of surgery for a gastric ulcer, for example—until now it has been cost increasing (see Lindgren & Lyttkens, in press).

It turns out that the effect of aging on health care expenditure is extremely difficult to foresee. First of all, health care has an income-elastic demand, meaning that as income goes up, so does demand for health care. Expenditure on health care closely follows a trend: The higher the GDP, the higher the proportion spent on health care. Thus, at least until now, demography is not the main explanation for increases in spending on health care. Adding costs for increasing quality at the same rate as real wage growth would increase health care costs by a factor of around 3.5% in addition to keeping per capita cost constant up to year 2040 (Lindgren & Lyttkens, in press).

Health care has a greater weight in older people's utility and demand. Aging means that older people's influence increases, both as consumers and via the political process as voters. We can thus expect that expenditures on health care will increase and do so more than just in response to demographic change. Furthermore, Richard Disney (2007) shows that aging will be associated with a larger welfare state.

### *Pensions*

In his seminal article, Paul Samuelson (1958) analyzes the problem of how to support oneself in old age. He uses an overlapping generations model with three coliving generations, two working and one retired (children are assumed to belong to the adults' families and are outside the analysis). To focus on the necessity of cooperation of generations, he assumes that nothing keeps; trade with nature is not possible, and there is no capital market. During the working periods, each person or generation produces one unit.

Samuelson (1958) shows that one way of solving the life cycle problem of smoothing the consumption possibilities, when working capacity does not suffice for all periods, is to make the two working generations pay a percentage of their incomes as a tax and use the revenue for payments to the old generation. One important conclusion from this setup of a contract between generations, a *pension system,* is that the rate of return on contributions paid equals the population growth rate in the economy. Remember that there is no capital in this economy and thus by assumption no productivity growth. With a positive population growth rate, the number of workers increases; with a fixed contribution rate, the sum of contributions increases, and so do the outgoing payments to the old.

This solves the life cycle problem. However, Samuelson concludes that this kind of system will not come into

existence because older persons do not have a negotiatory position to persuade the working generations to accept such a system, nor do they have the power to enforce it because they are in the minority. He points to the fact that life cycle redistribution is a linear business, not a circular one. That is, if A (old) borrows from B (middle-aged) and C (young), A will not be around to repay B in the next period, when B is old and needs support. The solution suggested by Samuelson is a social contact between generations; B gives part of his or her income to A, being convinced that in the period when B is old, C, now being middle-aged, and the new generation, D, now young, will do the same for him or her. In this simple setting, Samuelson shows the workings of a pay-as-you-go pension system and also that the rate of return in such a system equals the growth rate in the economy.

All industrialized countries have public pensions systems, often supplemented with pensions negotiated between the parties in the labor market and with private pensions. In most countries, public pensions provide the major part of income during retirement and constitute a large part of public sector expenditure. These expenditures are forecast to increase substantially with aging; hence, reforming the existing pension systems has become a top priority on political agendas.

Old age pensions can be organized in a number of ways. The main choices to make are the following:

- Public or private
- Obligatory or voluntary
- Pay-as-you-go or funded
- Defined benefit or defined contribution
- Basic or earnings related
- Redistribution or actuarial
- Indexed with prices, growth, or interest rate (rate of return in the capital market) during contribution years and during benefit receiving years.

The specific design chosen has different effects on the distribution between generations, on the rate of return the system gives, and also on how robust the system will be in response to demographic changes.

Public pensions are more often than not obligatory, meaning that the individual is not free to choose a life cycle consumption pattern according to his or her preferences. If the restriction imposed by the pension system is binding, this means a loss in utility. The higher the level of the system, the higher the probability of a binding restriction, which is probably more binding in low-income groups than in high-income groups. The reasons put forward to justify an obligatory system and this loss are the threat of free riders, the fear of myopia among younger people, and the desire to use the pension system for redistribution among socioeconomic groups.

A pay-as-you-go system may be described by its budget restriction:

$$q \times w \times L = b \times R, \tag{2}$$

where $q$ is the contribution rate, $w$ is the average wage, $L$ is the labor force, $b$ is the average benefit, and $R$ is the number of retirees. The left-hand side shows the sum of contributions (in a time period, a year, for example) and the right-hand side shows outgoing expenditures in that same period—that is, to contemporary retirees. The character of a pay-as-you-go system to be an implicit social contract between generations becomes evident with this formulation.

For the system to be in balance, the equality sign has to hold. Assume that wages, the labor force, and the number of retirees are exogenous to the pension system. It is then obvious that with demographic changes like those expected in the future, either the contribution rate has to be increased or the benefit level decreased, or both, in order to maintain a balanced system.

Rearranging equation 2 shows even more lucidly the importance of demography in a pay-as-you-go system:

$$q = b/w \times R/L, \tag{3}$$

where $b/w$ is the replacement rate—that is, the benefit in relation to wages—and $R/L$ is the dependency ratio.

Table 57.5 is to be read in the following way: With a dependency ratio, $R/L$, of 0.33, a contribution rate of 16.5% suffices for a replacement rate of 50%. If a higher replacement rate is wanted, for example 60%, then a contribution rate of 20% is needed. The combined information in Tables 57.3 and 57.5 leaves no doubt that aging will lead to pressure on pay-as-you-go pension systems.

The budget restriction shows the working of a pay-as-you-go system from society's point of view, but it can also be used to show it from an individual's life cycle perspective as well. With such a perspective, $L$ may be seen as the number of years the individual works, say from 20 to 64 years of age, giving 45 working years. $R$ is then the number of years in retirement, say from 65 to 80 years of age. Using these figures in equation 3 gives

$$q = b/w \times 15/45. \tag{4}$$

**Table 57.5** The Required Contribution Rate at Different Combinations (Assumptions) of the Replacement Ratio and the Old Age Dependency Ratio

| | *R/L 0.33*[a] | *R/L 0.40* | *R/L 0.50* | *R/L 0.60*[b] |
|---|---|---|---|---|
| *b/w* 50% | 16.5% | 20.0% | 25.0% | 30.0% |
| *b/w* 60% | 20.0% | 24.0% | 30.0% | 36.0% |

SOURCE: Author's calculations using equation 3.

(a) Corresponds to 45 years of work (ages 20 to 64) and 15 years as a retiree (ages 65 to 80).

(b) Compare the forecasts for Germany and Italy in 2050.

Assume that the individual wants a replacement rate of 60%. It is then obvious that the contribution rate has to be 20% (see also Table 57.5). The equation can be used to illustrate different changes. Assume, for example, that the individual wants to retire earlier, other things being equal, at 60 years of age. Then we get

$$q = b/w \times 20/40, \tag{5}$$

with a replacement rate, $b/w$, of 60%, the contribution rate has to be increased to 30%. Evidently early retirement is very expensive.

A last example here is increases in longevity. Assume an increase of 5 years, other things being equal. This gives us

$$q = b/w \times 20/45. \tag{6}$$

Again, with a $b/w$ of 60%, the contribution rate to balance the system is 27%.

The effects of aging on pay-as-you-go systems made policy makers and pension experts recommend reforms to change from pay-as-you-go systems to funded ones (see, e.g., World Bank, 1994).

In a funded system, the contributions paid during working years are put into a fund, which may invest in the stock market or hold equities, government bonds, or both. The fund increases each year with new contributions and with compound interest earned by the fund. At the time of retirement, the fund can be used for buying an annuity, the size of which depends of course on its value at the time of retirement.

There are pros and cons with either way of organizing a pension system. Obviously, the rate of return differs; in a pay-as-you-go system, it equals (approximately) the economic growth rate, and in a funded system, it equals the rate of return in the capital market. Historically, and in the long run, the rate of return in the capital market has surpassed the growth rate. This led Martin Feldstein (1974), among others, to the conclusion that funded systems were to be preferred; the same benefit would be possible with lower contribution rates. Thus, both aging and an expected discrepancy between growth and the interest rate seem to favor a funded system. Feldstein's conclusion builds, however, on the assumption that funding—and the eventual increased capital supply—will not affect the rate of return. If a great number of countries were to follow the advice of transforming their pay-as-you-go systems into funded ones, it is hard to believe that there would not be a downward pressure on the interest rate. Furthermore, it is worth noting that funded systems, to a certain extent, are also sensitive to demographic changes; for example, with a funded system, when baby boomers enter retirement and sell their funds to buy annuities and the increased capital supply meets a smaller working generation, the price—that is, the rate of return—will most certainly go down. Axel Börsch-Supan, Alexander Ludwig, and Joachim Winter (2006) show that demographic effect on the rate of return depends on the degree of funding as well as on the degree on international openness in the capital market.

It should be noted that both kinds of pension systems are exposed to risk; a pay-as-you-go system is exposed to the risk of a deteriorating economic performance and low growth, and a funded one is exposed to the risk of capital market crises. Thus, it may be a good idea to diversify, to use both devices.

Whether pay-as-you-go or funded, pensions are mainly life cycle savings with a smaller part being insurance. The insurance is in order to cover an extraordinarily long life—that is, the risk of outliving one's means. Insurance uses the law of large numbers, meaning that it is enough to have a buffer (savings) covering the average expected number of years as a retiree. With insurance, the risk of living longer than average life expectancy is pooled among the participants in the pension scheme. Because individuals have risk aversion, without insurance, the individuals would keep a too large buffer, which reduces expected utility.

The budget restriction of a pay-as-you-go system clarifies the difference between a *defined-benefit* (DB) and a *defined-contribution* (DC) system. In a DB system, the benefit ($b$) or the replacement rate ($b/w$) is fixed. Demographic and economic changes then have to be counterbalanced by changes in the contribution rate ($q$) in order to keep the system in balance. This means that the costs of adaptation to these changes fall on the working generation. As this discussion has shown, aging leads to a shrinking labor force; in a DB system, this requires raised contribution rates. In a DC system, the contribution rate is fixed and benefits are then determined by the sum of contributions divided by the number of retirees. Aging, in the form of increased life expectancy, means that yearly benefits decrease in proportion to the increase in the number of years as a retiree. In the last decade, there has been a trend away from DB to DC systems, indicating a shift in who bears the risk of increased life expectancy, from the working generation to the retired one (Whitehouse, 2007).

The natural index in a pay-as-you-go system is the growth rate (again, see Samuelson, 1958). However, many countries have chosen to index by prices, the reasons often stated to be economic. This implies, though, that the development of the pension system does not follow the development of the economy and will thus be exposed to economic changes because it does not adapt automatically. The standard of living of different generations will be determined by economic growth; with a high economic growth, the retirees' standard of living falls behind that of the working generation, and vice versa.

In an actuarial system, there is no redistribution ex ante. The sum of expected contributions equals the sum of expected benefits. Of course, there will be redistribution ex post; some people will end up as losers, living shorter than the expected average, and some as winners. If there are no

systematic differences between people or socioeconomic groups, then the system is actuarial. Note that if unisex life tables are used, there is redistribution in favor of women. In an actuarial system, with the contribution rate determined as a percentage of the wage, the system will be earnings related. Because women have a different labor market performance from men, with market work interrupted for child rearing, earnings-related systems are sometimes assumed to be disadvantageous to women. However, this turns out not to hold generally; see Ann-Charlotte Ståhlberg, Agneta Kruse, and Annika Sundén (2005) for an analysis of gender and the design of pension systems.

### *The Political Economy of Public Pensions and Aging*

Edgar Browning (1975) shows that in a democracy with majority voting, a pay-as-you-go system will expand beyond its optimal level. To show this, he uses an overlapping generations model with three generations, one young, one middle-aged, and one old. The young and middle-aged are working, earning $y_y$ and $y_m$ respectively. To introduce a pay-as-you-go system, they all vote on their most preferred tax or contribution level, knowing that the sum of contributions paid into the system is to be used for benefits to the contemporary old generation. The voters assume that the tax rate will not be changed in their lifetime.

Assume that the individuals maximize their lifetime utility. The old individual has only one remaining period left to live. This period is spent in retirement; the old person will not pay any contributions but will receive benefits. At the extreme—for example, neither taking the utility of children and grandchildren into account nor expecting negative incentive effects—the old person will vote for a tax and contribution rate of 100%, $t^{old}$. The young individual has his or her entire life ahead of him or her and will pay contributions during a whole working life and get benefits during retirement. The preferred contribution rate by the young will smooth the consumption path over the life cycle, maximizing utility. The young person's chosen contribution rate is thus the optimal one, $t^*$, as the entire life is taken into consideration. The middle-aged generation's most preferred tax rate will be between the old person's and the young person's. What the middle-aged generation has left is one period of working and contributing to the system and one period of getting benefits: $t^* < t^{middle\text{-}aged} < t^{old}$. With majority voting, the median voter casts the decisive vote, and the system will be greater than its optimal level. The outcome builds on rather simplified assumptions, for example, that the only aspect of voting is the level of the system and that voting takes place only once. Other researchers have qualified the Browning model and made assumptions closer to actual pension systems and more realistic voting procedures; even so, the Browning result holds well (see, e.g., Sjoblom, 1985).

This result is even reinforced by aging. Aging means that the age of the median voter increases. By the turn of this century, the age of the median voter was well below 50 years of age; in 35 years, it will increase to 54 in Germany, 55 in Italy and Spain, and 56 in France, to mention a few countries with rapidly aging populations (Galasso, 2006; Uebelmesser, 2004). Figure 57.2 shows how the median voter's considerations change with increased age.

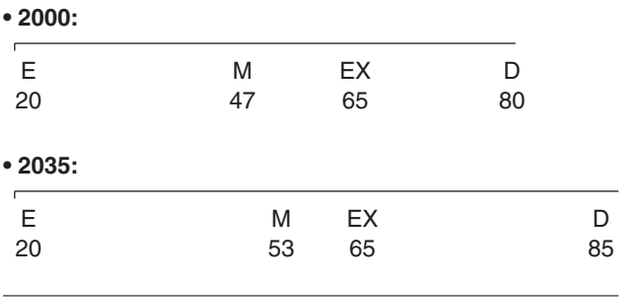


**Figure 57.2** Age of Median Voter and Life Expectancy: An Example

Assume that voting takes place in the year 2000 and that the median age is 47, retirement age 65, and expected remaining lifetime 15 years. Contributions before the age of 47 are sunk costs. The median voter knows that he or she has another 18 years of contributions (compared to the 40 to 45 years a person entering the labor market has) and 15 years of benefits. It is assumed that the median voter's age has increased 35 years later to 53 years, the retirement age is the same, but the expected lifetime has increased by 5 years. Voting in 2035, the median voter now has 12 years of contributions and is expected to get benefits during 20 years. A high contribution rate will certainly give a high payoff.

Aging gives rise to an opposite effect as well by exerting a downward pressure on the growth rate in the economy—that is, a downward pressure on the rate of return in a pay-as-you-go system. Therefore, it may be attractive to downscale the pay-as-you-go system and replace it with a funded one. So far, according to empirical results, the former tendency seems to have been the stronger:

> The most striking result . . . lies in the strong and significant positive effect of median voter age on program size. . . . one year adds half a percentage point to the GNP share of social security benefits. (Breyer & Craig, 1997, p. 719)

This also implies that reforming a pension system in response to the strain aging puts on the systems is politically difficult. Reform efforts have been met with fierce political resistance, especially from the so-called gray panthers. This notwithstanding, there has been a plethora of reforms in a number of countries where aging has put the pension systems under strain. The reforms range from marginal retrenching changes to radical parametric ones (Galasso 2006; Martin & Whitehouse, 2008). Radical reforms are of course more difficult to accomplish. Hans-Werner Sinn and Silke Uebelmesser (2002) analyze the

political possibilities of such a change in the German system and conclude that it will be possible until around 2015. After that date, Germany will turn into a gerontocracy, and such a change will not be politically feasible.

In Sweden, a radical reform was decided on in the middle of the 1990s. The system was changed from a DB to a DC system—the major part being kept as a pay-as-you-go system and a minor part transformed to a funded one—from price indexation to (in the main) wage indexation and with annuities depending on life expectancy. It is regarded as being a stable pension system despite the aging of the population (for a description see Kruse, in press; Palmer, 2006). Jan Selén and Ann-Charlotte Ståhlberg (2007) and Kruse conclude that the design of the transition rules was a crucial factor to form a majority in favor of the new system, together with high political competence. Other countries as well have implemented more or less similar reforms. Robert Holzmann and Edward Palmer (2006) thoroughly describe and analyze the pros and cons of these reforms.

## Possible Remedies and Concluding Remarks

There is no doubt, first, that aging will occur with increased old age dependency ratios, and second that aging will affect the extent of resources channeled to the elderly as well as affecting the possibilities of financing these resources. The prognoses give an increasing gap between available resources and future demands due to aging, ceteris paribus. Furthermore, aging is not the only foreseen challenge; in combination with ongoing globalization, aging is supposed to increase the strain. However, the ceteris paribus assumption will not come true. If there is anything to learn from history, it is that economies are flexible, adapting to new circumstances. The adaptation may, however, be costly.

To close the gap between the increasing demands due to an increased number of older persons and decreasing support possibilities due to a shrinking workforce, a number of measures aiming at both the demand and the supply side will probably be called for.

Immigration is often suggested as a remedy against the decreasing number of people of working age. However, demographers show convincingly that immigration is not a remedy; see for example Tommy Bengtsson and Kirk Scott (in press). Calculations show that the amount of immigration needed in order to compensate for demographic change and aging by far surpasses what is deemed possible to accommodate. Furthermore, even if the only accepted immigration would be people in working ages, immigrants also eventually get old and needy. Besides, such immigration assumes a rather odd immigration policy.

Another possibility would be to increase fertility, for example by stimulating fertility by subsidizing child rearing, making it possible to combine market work and children. Such policies are assumed to have kept fertility rates at rather high levels in France and Sweden, for example, in comparison with Germany, Italy, and Spain. However, fertility seems to be difficult both to predict and to influence. Besides, more children will at least initially not diminish the gap. On the contrary, it will reduce market labor supply because time is an important input into child rearing.

To raise taxes would be another way to close the gap. However, taxes give rise to incentive effects, which may reduce the tax base and cause deadweight losses. The deadweight loss increases with the square of the increase in the tax rate, making it an expensive way to go. There are, however, tendencies of a reformed tax structure, much in response to globalization, but it may also facilitate the financing of the welfare state (Bergh, 2008; Hansson, in press). Examples are pension reforms with a tighter connection between contributions and benefits, for example in the French and German point systems and in the Swedish defined-contribution system. The Japanese and German old age care insurance systems are other examples. Such earmarked taxes reduce the distortive effect of a tax.

An increased labor supply seems to be many countries' most favored measure to meet the strain of aging. For a long time, the opposite was common, with a number of countries taking measures to facilitate early retirement (see Gruber & Wise, 1999). To meet the challenges of aging, a number of countries are now trying to reverse these measures. Mandatory retirement ages are increased, and ways for early exit have been closed. Again, a closer link between taxes or contributions and benefits gives incentives for working longer hours and postponing retirement. Increased labor supply may be the solution to the financing problem caused by aging. Policy measures to give incentives in this direction are thus called for.

## References and Further Readings

Ando, A., & Modigliani, F. (1963). The "life cycle" hypothesis of saving: Aggregate implications and tests. *American Economic review, 53*(1), 55–84.

Bengtsson, T., & Scott, K. (in press). The aging population. In T. Bengtsson (Ed.), *Population aging and the welfare state.* New York: Springer-Verlag.

Bergh, A. (2008). A race to the bottom for the big welfare states? In A. Bergh & R. Höijer (Eds.), *Institutional competition* (pp. 182–201). Northampton, MA: Edward Elgar.

Börsch-Supan, A., Ludwig, A., & Winter, J. (2006). Aging, pension reform and capital flows: A multi-country simulation model. *Economica, 73,* 625–658.

Breyer, F., & Craig, B. (1997). Voting on social security: Evidence from OECD countries. *European Journal of Political Economy, 13,* 705–724.

Breyer, F., & Felder, S. (2006). Life expectancy and health care expenditures: A new calculation for Germany using the costs of dying. *Health Policy, 75,* 178–186.

Browning, E. K. (1975). Why the social insurance budget is too large in a democracy. *Economic Inquiry, 13,* 373–388.

Burtless, G. (2008, February). *The rising age at retirement in industrial countries* (CRR Working Paper No. 2008-6). Chestnut Hill, MA: Center for Retirement Research.

Cutler, D. M., Poterba, J. M., Sheiner, L. M., & Summers, L. H. (1990). An aging society: Opportunity or challenge? *Brookings Papers on Economic Activity, 21*(1), 1–73.

Disney, R. (2007). Population ageing and the size of the welfare state: Is there a puzzle to explain? *European Journal of Political Economy, 23,* 542–553.

Erlandsen, S., & Nymoen, R. (2008). Consumption and population age structure. *Journal of Population Economics, 21,* 505–520.

EU Economic Policy Committee. (2003, October). *The impact of aging populations on public finances: Overview of analysis carried out at EU level and proposals for a future work programme* (EPC/ECFIN/435/03 final). Brussels, Belgium: Author.

Feldstein, M. (1974). Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy, 82*(5), 905–926.

Galasso, V. (2006). *The political future of social security in aging societies.* Cambridge: MIT Press.

Gruber, J., & Wise, D. A. (Eds.). (1999). *Social security and retirement around the world.* Cambridge, MA: National Bureau of Economic Research.

Hansson, Å. (in press). In this world nothing is certain but death and taxes: Financing the elderly. In T. Bengtsson (Ed.), *Population aging and the welfare state.* New York: Springer-Verlag.

Holzmann, R., & Palmer, E. (Eds.). (2006). *Pension reform: Issues and prospects for non-financial defined contribution (NDC) schemes.* Washington, DC: World Bank.

Kruse, A. (1997, February). Replacement rates and age specific public expenditure. In European Commission (Ed.), *Pension systems and reforms: Britain, Hungary, Italy, Poland, Sweden* (P95-2139-R, pp. 191–209). Budapest, Hungary: Phare ACE Programme.

Kruse, A. (in press). A stable pension system: The eighth wonder. In T. Bengtsson (Ed.), *Population aging and the welfare state.* New York: Springer-Verlag.

Lindgren, B., & Lyttkens, C. H. (in press). Financing health care: A Gordian knot waiting to be cut. In T. Bengtsson (Ed.). *Population aging and the welfare state.* New York: Springer-Verlag.

Martin, J. P., & Whitehouse, E. (2008, July). *Reforming retirement-income systems: Lessons from recent experiences of OECD countries* (OECD Social, Employment and Migration Working Paper No. 66). Paris: Organisation for Economic Co-operation and Development, Directorate for Employment, Labour and Social Affairs.

Organisation for Economic Co-operation and Development. (2002). *Policies for an aging population: Recent measures and areas for further reforms* (Macroeconomic and Structural Policy Analysis Working Paper No. 1). Paris: Author.

Organisation for Economic Co-operation and Development. (2008). *Employment outlook 2008.* Retrieved December 20, 2009, from http://www.oecd.org/document/25/0,3343,en_2649_33927_40762969_1_1_1_37457,00.html

Palmer, E. (2006). What is NDC? In R. Holzmann & E. Palmer (Eds.), *Pension reform: Issues and prospects for non-financial defined contribution (NDC) schemes* (pp. 17–33). Washington, DC: World Bank.

Samuelson, P. (1958). An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy, 66*(6), 467–482.

Selén, J., & Ståhlberg, A.-C. (2007). Why Sweden's pension reform was able to be successfully implemented. *European Journal of Political Economy, 23,* 1175–1184.

Sinn, H.-W., & Uebelmesser, S. (2002). Pensions and the path to gerontocracy in Germany. *European Journal of Political Economy, 19,* 153–158.

Sjoblom, K. (1985). Voting for social security. *Public Choice, 45,* 227–240.

Sjögren Lindquist, G., & Wadensjö, E. (in press). The labour market for older workers in Sweden: Changes and prospects. *European Papers on the New Welfare.*

Söderström, L. (2008). *The economics of social protection.* Northampton, MA: Edward Elgar.

Ståhlberg, A.-C., Kruse, A., & Sundén, A. (2005). Pension design and gender. *European Journal of Social Security, 7*(1), 57–79.

Uebelmesser, S. (2004). Political feasibility of pension reforms. *Topics in Economic Analysis and Policy, 4,* 20.

United Nations. (2007). *World population prospects: The 2006 revision population database.* Retrieved December 20, 2009, from http://ecoglobe.ch/population/e/p2k07407.htm

Whitehouse, E. (2007, October). *Life-expectancy risk and pensions: Who bears the burden?* (OECD Social, Employment and Migration Working Paper No. 60). Paris: Organisation for Economic Co-operation and Development.

World Bank. (1994). *Averting the old age crisis: Policies to protect the old and promote growth.* Washington, DC: Author; Oxford, UK: Oxford University Press.

# 58

# Agricultural Economics

Patricia A. Duffy
*Auburn University*

Agricultural economics is an applied field of economics that focuses primarily on food and fiber production and consumption. Defining the boundaries of agricultural economics can be difficult, however, because issues outside these traditional areas have become increasingly important to the profession in recent years. Agricultural economists engage in work ranging from farm-level cost accounting to assessing the consumer impact of food safety and nutrition labeling to analyzing worldwide agricultural trade patterns and a host of other real-world issues. Accordingly, the Agricultural and Applied Economics Association (AAEA), the largest professional organization for agricultural economists in the United States, currently recognizes a variety of topic areas under the broad disciplinary umbrella, including community and rural development, food safety and nutrition, international trade, natural resources and environmental economics, consumer and household economics, markets and competition, agribusiness management, and production economics (AAEA, n.d.).

Because many other topics in applied economics, including natural resources and environmental economics, are handled separately in this volume, this chapter concentrates on the field of agricultural economics as it relates to the food and fiber sector. It is worthwhile to note, however, that within the profession, an increasingly large share of university and government agricultural economists focus their teaching, research, and outreach efforts on work outside this sector.

The broadening of the field of agricultural economics reflects changes in the national economy. Dramatic changes in agricultural technology over the course of the twentieth century have meant that fewer and fewer people have been needed to feed more and more people. According to the National Agricultural Statistics Service of the U.S. Department of Agriculture, production agriculture (e.g., farming) directly employs fewer than 3 million workers, including farm owners and operators, or about 1% to 1.5% of the total U.S. workforce, depending on how the workforce is measured. Even as far back as 1946, the Nobel Prize–winning agricultural economist Ted Schultz (1946) commented on the decreasing need for labor in production agriculture and the declining role of production agriculture in U.S. employment. However, the food and fiber system, which includes processing, agricultural input production, and retail sales of food and fiber products remains one of the largest sectors in the U.S. economy, accounting for about 12% of the total U.S. gross domestic product, and employing about 16% of U.S. workers (Edmondson, 2004.) Thus, agricultural economics addresses issues of importance to a variety of clientele groups in the United States, including but not limited to those directly involved in farming. As one popular bumper sticker phrased it, "If you eat, you are involved in agriculture."

Further, although production agriculture may not hold a large, direct share of the U.S. economy, in many developing nations, agriculture remains the largest employer, and trade in agricultural products is an important engine for economic growth. Organizations focusing on international development, such as the World Bank, the Food and Agricultural Organization of the United Nations, and the International Food Policy Research Institute are all strong employers of agricultural economists. The International Association of Agricultural Economists (IAAE, 2008) provides a worldwide network for the discipline, and professional societies for agricultural economists can be found in countries around the world. The wide scope of the field and its overall importance to both the U.S. economy and

worldwide economic systems thus warrants the inclusion of this chapter in *21st Century Economics: A Reference Handbook.*

In this chapter, an overview of the field of agricultural economics is provided, focusing on the applications and contributions affecting the food and fiber sector. The chapter begins with a brief history of the field. Next, the underlying theoretical underpinnings are discussed. This theoretical section is followed by an outline of quantitative tools used in agricultural economics, then by short descriptions of a few major subfields of the discipline. Some notable agricultural economists and their contributions to the profession are then discussed, followed by a section on the future of the profession. Finally, this chapter concludes with a list of references and some suggested further readings for those wishing to gain a deeper knowledge of this topic.

## Short History of Agricultural Economics

The field of agricultural economics traces its origins largely to the growing demand at the turn of the twentieth century for college-educated professionals who could address the special concerns of the agricultural sector. A seminal textbook, *Farm Management* by Cornell agricultural economist George F. Warren, was published in 1913. Shortly after, in 1919, the *Journal of Farm Economics* was launched by the American Farm Economics Association with the stated purpose of serving those interested in "economic forces and influences as they operate to affect the business of farming," ("Forward," 1919, p. 1). At the time the journal was launched, a sizable proportion (over 20%) of the population lived on farms. From the end of the Civil War to the mid-1930s, with the westward movement of population, farm numbers rose sharply from about 2 million to a peak of nearly 7 million. Farming and the related concerns of rural residents were of widespread interest in the early decades of the twentieth century, and a journal dedicated to farm economics would have had a strong appeal at this time.

The journal continues to operate to this day; however, in 1967, its name was changed to the *American Journal of Agricultural Economics,* reflecting the broadening of interests that occurred over that time. The association that sponsored the original journal also continues to this day, although its name has twice changed, once to the American Agricultural Economics Association, and more recently to the Agricultural and Applied Economics Association, again reflecting the increasing scope of activities for its members.

The premiere issue of the *Journal of Farm Economics* included a presidential address by G. A. Billings of the U.S. government's Office of Farm Management, an organization established in 1905 that later gave way to the Office of Farm Management and Farm Economics and, finally, in 1961, to the present-day Economic Research Service of the U.S. Department of Agriculture. In addition, the issue contained an article on farm labor outlook by Secretary of Agriculture G. I. Christie and an article by H. W. Hawthorne, also from the Office of Farm Management, on information obtained from farm surveys. Notable to a modern-day agricultural economist is that, even from its inception long before the era of high-speed computing, the field was data driven. Also, a focus on practical, real-word problems is clearly evident in these historical documents.

At around the same period, in the early years of the twentieth century, agricultural economics (often then called *farm economics* or *farm management*) became an area of study in universities across the nation. Land-grant universities, in particular, with their focus on agricultural and mechanical arts, were fertile grounds for the development of this field. At the Agricultural and Mechanical College of North Carolina at Raleigh (which later became North Carolina State University), for example, a course in agricultural economics was required of all students in the College of Agriculture as far back as 1897, although a separate department, at the time called Agricultural Administration, was not established until 1923 (Bishop & Hoover, n.d.). At Michigan State University, a course in farm management was offered for the first time in 1906. At the University of Minnesota, a department for agricultural economics was established in the College of Agriculture in 1909 (Shaars, 1972). At Auburn University, in Alabama, a department of agricultural economics was established in the College of Agriculture in 1928 (Yeager & Stevenson, 2000). Interested readers are also referred to Bernard Stanton's (2001) history of agricultural economics at Cornell and Willard Cochrane's (1983) history of the discipline at the University of Minnesota.

In the first half of the twentieth century, farm-level issues remained the primary concern of research, outreach, and teaching efforts. One important area of endeavor was calculation of the cost of production for agricultural commodities. Other work involved agricultural marketing, credit, and then, in the years following the Agricultural Adjustment Act in 1933, the increasingly important subfield of agricultural policy. A glance at the table of contents of the *Journal of Farm Economics* will show, however, that even as far back as the 1930s, there was widespread interest in international trade in agricultural products. The scientific study of factors affecting consumer demand for agricultural products also dates to this period.

The period following World War II saw an expansion of the agricultural economics field with universities increasing the size of their faculties and the government employing larger numbers of agricultural economists. Membership in the American Agricultural Economics Association peaked in the 1980s and has been declining since.

The decades after World War II saw a proliferation of professional associations and sponsored journals in which agricultural economists could publish an expanding

array of papers related to research, teaching, and outreach. Regional associations differentiated themselves, to some degree, in terms of their concerns for agricultural problems specific to their geographic areas, based on either different agricultural practices and outputs or different regional institutions. In 1969, the Southern Agricultural Economics Association launched the *Southern Journal of Agricultural Economics.* This journal was renamed the *Journal of Agricultural and Applied Economics* in 1993, reflecting a trend away from regional identification and an increased emphasis on applied economic analysis outside the traditional agricultural sector. The Western Agricultural Economics Association began publishing the *Western Journal of Agricultural Economics* in 1977. From the outset, this journal specifically solicited articles on natural resource economics, along with those related to human resources, and rural development. In 1992, this journal changed its name to the *Journal of Agricultural and Resource Economics,* to emphasize its continued focus on the economics of natural resources and to remove its regional focus. In the same mode, the *Northeastern Journal of Agricultural Economics,* whose publication dates from 1984, was renamed in 1993 as the *Agricultural and Resource Economics Review.* The *North Central Journal of Agricultural Economics,* published from 1979 to 1990, was renamed as the *Review of Agricultural Economics* and then again renamed in 2010 as *Applied Economic Perspectives and Policy.* It is no longer associated with an individual regional association.

The number of international journals devoted to agricultural economics also expanded over a similar period. Currently, international journals include the *Canadian Journal of Agricultural Economics,* the *European Review of Agricultural Economics,* the *Australian Journal of Agricultural and Resource Economics,* the *ICFAI University Journal of Agricultural Economics* (India), and the *Journal of Agricultural Economics* (United Kingdom). The International Association of Agricultural Economists publishes *Agricultural Economics,* which was launched in 1986. The importance of the field internationally is also evidenced by the recent launch of the new journal, *China Agricultural Economics Review,* by faculty at the China Agricultural University in Beijing, China.

Another important change that has taken place over time involves the use of quantitative techniques. Although statistical and quantitative analyses were important in agricultural economics since the field's inception, the advent of computer technology has greatly expanded the methods that can be employed for the analysis of agricultural economic problems. Today, agricultural economists employ highly sophisticated statistical techniques, and most are adept at the use of a variety of computer software packages, from electronic spreadsheet programs to dedicated statistical software. Further discussion of quantitative techniques in agricultural economics follows later in this chapter.

## Undergraduate Education

Education in agricultural economics has evolved considerably from its early twentieth-century roots in farm management. A search of the College Board (n.d.) Web site yielded a list of over 100 universities in the United States and Canada offering 4-year degrees in agricultural economics or the closely related major of agricultural business.

In addition to university-wide general educational requirements, students majoring in agricultural economics or agricultural business typically take foundations courses in business, usually in economics and accounting, as well as a sampling of courses in agricultural sciences, such as agronomy, horticulture, and animal sciences. Courses in the home department cover topics such as agricultural marketing, agricultural finance, agribusiness or farm management, natural resource and environmental economics, and policy and trade. Additionally, most universities require quantitative courses, such as calculus and statistics, for all undergraduate majors in this field.

## Theoretical Underpinnings

As an applied field, agricultural economics combines basic theory, quantitative techniques, and institutional knowledge to explain or predict real-world phenomena. Thus, most agricultural economists use economic theory for generating hypotheses to be tested or as the basis for formulating a statistical model to be estimated. Microeconomic theory is the branch of theory most often used, although a few agricultural economists have applied theory from macroeconomics (see, e.g., Penson & Hughes, 1979). In addition to drawing heavily from the field of economics for its general theory and disciplinary home, agricultural economics also may draw on other business disciplines including finance, marketing, and management, depending on the problem at hand, although even in these cases, economic theory forms the foundation. Agricultural economics may also draw heavily on biologically based disciplines, engineering, or ecology. In Part II of this handbook, economic theory is discussed in considerable detail; hence, the material will not be repeated at length here. Rather, this section will instead focus on how agricultural economists use microeconomic theory in their applied work.

A theoretical economic model is an abstraction from reality. To be useful, it cannot be so simple as to ignore key interrelationships, but neither can it be so complex that it obscures these relationships. To formulate a good theoretical model, agricultural economists must be well versed in general economic theory and must understand the processes and institutions involved in the problem of interest. The best theoretical model in the world, for example, would be useless to a production economist who did not know the growing seasons or climate zones for the agricultural products of interest in the research problem.

When existing theory has not been sufficiently developed to provide necessary insights into an applied problem, agricultural economists have contributed to the theory in important ways. Two classic examples are Clark Edwards's (1959) work on asset fixity and supply and a work by James Houck (1964) on the relationship between the demand elasticity for joint products and the demand elasticity of the underlying commodity. Similarly, Zvi Griliches (1957) contributed greatly to the understanding of causes and consequences of technical change, beginning with his seminal paper on hybrid corn. A recent example of contribution to economic theory is the paper by Carlos Carpio, Michael Wohlgenant, and Charles Safley (2008), providing a framework in consumer decision making for distinguishing the effect of time as a resource constraint and time that provides enjoyment. Another recent paper extending economic theory is the Jeffrey LaFrance and Rulon Pope (2008) investigation of the homogeneity properties of supply functions in the Gorman class.

## Production Theory

Production theory is used in farm management applications, estimations of productivity growth, and formulation of estimates of the response of agricultural output to changes in economic conditions. The basis of production theory in microeconomics is the production function, a systematic way of showing the relationship between inputs and outputs, in physical terms. To produce an agricultural output such as corn, for example, land, fertilizer, labor, and other inputs are required. Although some may view the estimation of production relationships as the work of the biologically based fields, agricultural economists were among the pioneers in this area. Articles dating from the 1940s in the *Journal of Farm Economics* tackle the estimation of physical production relationships (Heady, 1946; Tintner & Brownlee, 1944).

An important principle related to the production function is the law of diminishing marginal returns (often called just the *law of diminishing returns*). The engineering law states that as additional units of a variable input are used in combination with one or more fixed inputs, the amount of additional output per additional unit of variable input will begin to decline. The law of diminishing returns can be traced back to the concerns of early economists such as von Thünen, Turgot, Malthus, and Ricardo and was one of the first principles addressed by agricultural economists (see, e.g., Spillman, 1923). It remains an important concept to this day, especially in the area of farm management.

Agricultural economists have long been interested in estimating the supply curve for agricultural products. Neoclassical production theory maintains that firm behavior is characterized by the maximization of profits subject to a set of technical and institutional constraints. In the basic perfectly competitive model, where perfect knowledge is assumed and producers are said to be price takers (i.e., no individual producer can influence market price for either output or inputs), the profit function can be used to solve mathematically for quantities supplied. Without requiring the assumption of perfectly competitive output markets, an alternative specification involves minimizing the cost of producing a given amount of output.

Because prices are fairly transparent in the market, in recent years, applied economists have typically started with indirect profit or cost functions (e.g., functions specified in terms of prices and costs) and used these to generate the functional form of their subsequent estimations of output supply or input demand. Readers interested in the properties of the functional forms specified for the production function and how they relate to indirect profit or cost functions and the derived supply function or input demand equations are referred to Robert Chambers (1988) or to Bruce Beattie and C. Robert Taylor (1985). For a classic example of the use of an indirect profit function to specify input demand and output supply equations, readers may see C. Richard Shumway's (1983) article on supply response of Texas field crops. Readers are also referred to Christopher O'Donnell, Richard Shumway, and V. Eldon Ball (1999) for a discussion of using flexible functional forms to specify input demand equations.

Although the perfectly competitive model has been the basis of much useful work, modifications to this model have been made to address real-world issues. For example, many factors important to agricultural producers, such as the output price to be received in the future, are uncertain. Thus, the role of producers' price expectations and their effect on supply response was explored in a seminal work by Marc Nerlove (1956). Additionally, output is not perfectly predictable because factors outside the producers' control, such as weather or pests, may affect it. The basic profit-maximization decision model may thus be altered to reflect producers' aversion to highly risky enterprises. For an overview of how risk is incorporated into agricultural decision making, readers may see Richard Just and Rulon Pope (2002) or J. Brian Hardaker, Ruud Huirne, Jock Anderson, and Gudbrand Lien (2004).

In field crop supply estimation, another important modification is the inclusion of variables to account for farm program provisions, such as support prices or acreage reduction programs, which may influence producer decisions. A method of accounting for government program effects on supply was developed by James Houck and Mary Ryan (1972). A theoretical framework to consider both farm program provisions and the impact of risk on agricultural supply for field crops was developed by Jean-Paul Chavas and Matthew Holt (1990) as the foundation for the empirical estimation of the aggregate response of soybean and corn acreage to changes in price, policy provisions, and risk.

The pages of agricultural economics journals will provide many other examples of the use of production economic theory to formulate models for empirical work. In addition to the types of work already considered, production economists have also provided insight on innovation

and structural change (see, e.g., Huang & Sexton, 1996) and the impact of research expenditures on productivity. Readers interested in a more advanced and extensive treatment of the work of agricultural economists in production economics are referred to the *Handbook of Agricultural Economics: Volume 1A: Agricultural Production* (Gardner & Rausser, 2001a).

## Consumer Theory

Consumer theory is also used extensively in agricultural economics, notably in the formulation of market demand curves for agricultural products or in applied price analysis. The basic behavioral hypothesis is that the consumer, in deciding among the myriad of products to buy, selects that combination and quantity that maximizes his or her utility (happiness) subject to a budget constraint. The budget constraint includes the consumer's income and market prices, both of which are assumed to be beyond the immediate control of the consumer. Income may also be assumed fixed for analytical convenience and to focus on the main objects of choice—namely, the quantities of the various goods the consumer purchases.

In some instances of demand equation estimation, a specific functional form for the utility function is specified, as in the almost ideal demand system model (Deaton & Muellbauer, 1980). More often, in estimation of demand for agricultural products, utility theory is used to place restrictions on estimated elasticities and to reduce the number of price variables included in the model through the maintained hypotheses of weak separability and two-stage budgeting.

A difference between agricultural economics and general economics, in terms of applications of demand theory, is that basic agricultural products are more likely to be homogenous or largely indistinguishable from each other than are other products in the market. Thus, a generalized demand curve for eggs, milk, or catfish filets makes more intuitive sense than a generalized demand curve for automobiles, which are distinguished by brand name and other features. Outside the agricultural sector, there are far fewer examples of truly homogenous products, other than raw minerals or metal ore. Because many agricultural products are homogeneous and the farm firms producing them have no individual control over the output price, generic advertising and promotion programs have been employed by the industries. Agricultural economists have provided valuable information to these industries on the effect of these promotion efforts both in U.S. markets and in export markets (see, e.g., Forker & Ward, 1993; Nichols, Kinnucan, & Ackerman, 1990).

Even within the agricultural sector, the assumption of homogenous products may not apply. Sales of agricultural land, for example, would not fit the assumption of a homogenous product. Many factors, including location, would distinguish one parcel from another. In addition, the market for land is usually thin, so that individual buyers and sellers may influence price. Alternative theoretical models have been applied in situations when the assumptions of perfect competition do not fit. Hedonic pricing models, which relate the price of a product to its attributes, have been applied to products that are not homogenous, one of the earliest applications in agricultural economics being by Frederick Waugh (1928) for vegetables. In markets for land, methods to account for spatial factors have also been applied. Agricultural economists have also incorporated theory related to asymmetric information into their conceptual models and contributed to the development of this vein of work. Even for products typically considered homogeneous, such as wheat or cotton, distinctions may be made in international trade models with respect to country of origin.

Demand theory applied in agricultural economics has incorporated the links between the farm and retail sectors. Exploration of the factors affecting marketing margins, or the farm-to-retail price spread, has been a fruitful area of applied research. A notable example of a paper that contributed to economic theory in this area as well as provided useful empirical information is by Michael Wohlgenant (1989).

Finally, because prices are determined by the intersection of supply and demand, agricultural price analysis is a subfield that draws from both consumer and producer theory. Agricultural economists working under this general umbrella have made strong contributions to the analysis of futures markets, the economics of storage, the economics of food labeling and food safety, and spatial price analysis. A reader interested in a more extensive treatment of the applications of consumer theory and price analysis by agricultural economists is referred to the *Handbook of Agricultural Economics: Volume 1B: Marketing, Distribution and Consumers* (Gardner & Rausser, 2001b).

## Equilibrium Displacement Models

One final concept is worth noting in this section. Elasticity of supply or demand relates the percentage change in a commodity's own price to the percentage change in the quantity supplied or demanded. Elasticities of demand and supply as well as price-transmission elasticities taken from previously published work can be used in applied analyses. These studies, often called *equilibrium displacement models*, trace the effects of a shock or shifter of either the demand or supply curve (or both) on market prices, quantities, and producer and consumer welfare. Complicated real-world linkages between retail and wholesale markets can be specified, with various degrees of market power assumed at different levels in the marketing chain. However, to provide good estimates of the likely effects of the shock, the elasticity estimates used in the equilibrium displacement model must be accurate. Hence, the professional standards for publication of supply and demand estimates, from which elasticities can be drawn, require that the researchers present both sound theoretical justification for the functional form of the model and employ appropriate quantitative methods to avoid introducing bias in the parameters upon estimation.

(For further discussion of the use of equilibrium displacement models in applied research, see Piggott, 1992.)

## Quantitative Techniques in Agricultural Economics

Agricultural economics makes heavy use of quantitative methods. Statistical techniques used by agricultural economics generally fall under the heading of *econometrics,* discussed in Chapter 5 of this handbook. As such, they will not be presented in great detail here. In observational data, as opposed to experimental data, many things change at once. Statistical methods of overcoming the limitations of the available data have been employed by agricultural economists for decades. Most published articles by agricultural economists involve the use of sophisticated econometric techniques. An excellent, although fairly technical, discussion of the use of and problems with these techniques was provided by Just (2008) in his presidential address to the AAEA.

### Operations Research Methods

Although not as widely used as econometrics, operations research techniques have been employed in agricultural economics almost since the development of these tools in the period during and following World War II. Techniques typically used by agricultural economists are of two general types: mathematical programming models and simulation models. Mathematical programming models have an objective function to be maximized or minimized, and simulation models generally do not.

Linear programming, one of the best-known operations research techniques, is widely used in farm management work and is often taught in undergraduate classes (see, e.g., Kay, Edwards, & Duffy, 2008, chap. 12). Linear programming models have a linear objective function to be maximized or minimized subject to linear inequality constraints. They can be useful models for firm-level analysis in agriculture, either in terms of profit maximization for the farm or in terms of calculating a least-cost diet for feeding livestock.

Even before the simplex method for solving linear programming models was formally developed by Dantzig in 1947, George Stigler (1945) investigated the problem of the minimum cost diet for human subsistence using approximation to solve the constrained minimization problem. The objective of the model in this paper was to minimize the cost of feeding a person, subject to keeping nutrient intake above or below certain levels necessary for health. Calculation of the thrifty food plan for the Food Stamp Program is a similar modern-day endeavor, although information on human nutrition and techniques to solve this sort of problem have improved vastly since Stigler's time.

Following the introduction of the simplex method of solution, linear programming was adopted rapidly in the agricultural economics profession so that by 1960, it was already well established (Eisgruber & Reisch, 1961). Earl Heady and Wilfred Candler (1958) contributed an influential and widely used textbook on this new technique, specifically geared toward uses in agriculture. Although most of the early applications were to farm management, the technique was also used for marketing and other areas.

With the advent of high-speed computers, feasible operations research methods in agricultural economics became more sophisticated. R. C. Agrawal and Earl Heady (1972), in an updated textbook, listed several extensions to the linear model, such as variable resource programming, integer programming, quadratic programming, nonlinear programming, and dynamic programming. Loren Tauer (1983) introduced Target MOTAD, an extension of linear programming that allowed for a theoretically consistent method of incorporating a producer's risk preferences into the decision framework.

Simulation, another operations research tool, also has early roots in the agricultural economics profession, tracing back at least as far as an article in the *Journal of Farm Economics* by Pinhas Zusman and Amotz Amiad (1965). An overview of early uses for simulation in agricultural economics was provided by Anderson (1974). Simulation models can be used at the firm level, to simulate, for example, the probable effects on output and income of alternative farm program proposals, or at the sector level, to simulate the probable impacts of technical change or changes in macroeconomic conditions. Models can incorporate real-world linkages and the probabilistic nature of some outcomes, such as weather events, farm yields, or prices. Some simulation models use econometric estimates as their core, and others are constructed from systematic observation of the important relationships.

## Areas of Concentration

In a recent paper presented to the AAEA and subsequently published in *Applied Economic Perspectives and Policy*, Gregory Perry (2010) pointed out the proliferation in areas of concentration in the association, from 12 in 1966 to more than 80 today, and the decline in the primacy of the traditional fields of farm management and marketing. Of approximately 2400 members of the association listed in the current directory, only 114 chose farm management as one of their areas of specialization, and 141 members chose marketing. One traditional area, agricultural policy, retains a fair share of the membership, with 291 members listing it as one of their subfields.

Space does not permit a discussion of each area of concentration in agricultural economics, and many of these, such as diet, consumption, and health, overlap with other fields of applied economics. Hence, the following paragraphs are

devoted to providing a quick summary of a few agriculturally focused subfields. Readers interested in a broader look at the work of agricultural economics are referred to John Penson, Oral Capps, C. Parr Roson, and Richard Woodward (2006).

## Farm Management and Production Economics

Although these are separate areas of concentration, there is considerable overlap between the two. Farm management is the area in agricultural economics primarily concerned with the profitability of a farm firm. Farm management professionals develop both cost-of-production estimates (based on already realized outcomes) and planning budgets (projections for future outcomes) for various farm enterprises, such as cotton production or a cow-calf operation. They are concerned with topics such as optimal machinery size for a given farm, risk management, income tax management, and a host of other issues that directly or indirectly affect the profits of the farm. Because farm policies and natural resource protection policies affect farm profitability, there can be considerable overlap between the work of farm management professionals and those specializing in policy or natural resource and environmental economics.

Production economics has its roots in farm management. Factors that make an individual farm profitable or not profitable will also affect the overall supply of a commodity. Production economists are concerned with such topics as the impact of technical change on agricultural output, efficiency gains in the use of inputs, returns to agricultural research, and the aggregate impact in terms of output of farm and environmental policies. Farm management specialists and agricultural production economists will often engage in interdisciplinary or multidisciplinary work with production scientists, such as agronomists or animal scientists. It is not unusual for researchers in this field to coauthor work in the journals of these other disciplines. The *Agronomy Journal,* for example, lists economics as one of its target areas, and there are currently over 100 papers in the area of economics posted on this journal's Web site.

## Agricultural Marketing and Prices

Agricultural marketing can involve any of the processes that move an agricultural commodity from the farm gate to the dinner plate. A specialist in agricultural marketing may direct efforts toward helping establish a farmers' market, for example, or study the relationship between the futures price of a commodity and its cash price. Food processing and distribution are also areas of interest inside this concentration, and in this respect, agricultural marketing can have considerable overlap with the agribusiness area. Although agricultural marketing has much in common with business marketing, agricultural marketing distinguishes itself in terms of a heavier reliance on microeconomic theory, a greater focus on homogenous products, and an emphasis on the marketing institutions related to agriculture, such as forward contracting and the commodity futures and options market. Interested readers are referred to Richard Kohls, Joseph Uhl, and Chris Hurt (2007).

The subfield of agricultural prices examines price determination for agricultural products. As such, it is concerned with the intersection of supply and demand. Determination of marketing margins and the impact of industry structure on price are topics of investigation in this concentration, as is price differentiation based on spatial or quality factors. Readers interested in learning more about agricultural price analysis are referred to William Tomek and Kenneth Robinson (2003) or John Goodwin (1994).

## Agricultural Policy

Since the inception of agricultural policy in the 1930s, agricultural economists have been at the forefront of this focus area. The cost of the major provisions of the U.S. farm bill averaged more than $45 billion per year over the 2002 to 2007 period, including $15 billion in farm support payments and more than $29 billion in Food Stamp Program costs (Chite, 2008). Given the sizable outlays, it is not surprising that agricultural economists have been interested in the impact of these programs. Analysis of farm programs may involve their effect on farm profitability, farm structure, or choice of inputs and outputs. Readers interested in learning more about agricultural policy are referred to Ronald Knutson, J. B. Penn, and Barry Flinchbaugh (2007).

## Agribusiness Management

The term *agribusiness* can be applied to any firm involved in the food and fiber industry, from a farm to a retail store or restaurant to an input supplier, such as a firm that manufactures fertilizer or tractors, to an agricultural service provider, such as a veterinary operation or an agricultural lender. As such, this area overlaps or subsumes several other areas. As a research category within agricultural economics, the term generally refers to studies dealing with input suppliers, processors, retailers, or other sectors beyond the farm.

Perry (2010) pointed out that a shift has occurred in undergraduate education, with a decline in the number of undergraduate degrees in agricultural economics since the early 1990s, largely offset by an increase in the number of degrees in agribusiness. Distinguishing the two degree programs, in terms of coverage, is quite difficult, and some programs that offer a degree in agricultural economics have an agribusiness option. Further complicating a clear definition of the term *agribusiness* is its use in other departments inside colleges of agriculture as an option within, for example, agronomy or animal science. Readers

interested in learning more about agribusiness management are referred to Steven Erickson, Jay Taylor Akridge, Fred Barnard, and W. David Downey (2002).

### Agricultural Finance

Agricultural finance is the subfield of agricultural economics that deals most explicitly with capital and the use of credit. Topics in this subfield would include such endeavors as investment analysis, the effects of debt load on farm survival and structure, taxation, capitalization, and interest rates. The farm credit system was established in 1916 as a source of funds for agriculture, and some of the work in agricultural finance has focused on this system and its impact on the agricultural sector. The line between agricultural finance and other subfields is not always easy to draw, however, because investment and the use of credit are important in farm management and agricultural marketing. Considerable work on risk and uncertainty in agriculture has originated from research focused on agriculture finance. A journal dedicated to agricultural finance, the *Agricultural Finance Review,* began publication in 1938 and continues to this day. For more information about agricultural finance, see Peter Barry, Paul Ellinger, John Hopkin, and C. B. Baker (2000).

### Other Areas

Agricultural economists have made considerable contributions to development economics, with Ted Schultz, a notable agricultural economist, winning a Nobel Prize for his work in this area. Many agricultural economists specialize in resources or environmental economics. Others focus their work in the area of international trade. In recent years, agricultural economists have made contributions to the area of health economics, particularly concerning the link between diet and health. Given the large outlays from the U.S. Department of Agriculture for nutrition support programs, especially the Food Stamp Program, some agricultural economists have conducted studies to estimate the impact of these programs on the nutritional status or food security level of the recipients. In all of these endeavors, the line between agricultural economics, general applied economics, and consumer economics is almost impossible to draw.

## A Handful of Notable Agricultural Economists

Those interested in a thorough study of the contributions of agricultural economists to the solution of applied problems affecting agriculture and related areas are referred to the list of fellows of the AAEA (n.d.). The Association began naming fellows in 1957, and short biographies of the fellows, including their major contributions to the field, are published in the December issues of the journal. Because of space limitations, only a handful of these notable agricultural economists can be discussed in this chapter.

Theodore Schultz (1964) received a Nobel Memorial Prize in Economics in 1979 in recognition for his work in development economics. As he pointed out in his acceptance speech, the majority of the world's people are poor, and most of the poor are involved in agriculture. He was also one of the pioneers in exploring human capital and its role in development. A slim volume, *Transforming Traditional Agriculture*, is an excellent starting point for understanding Ted Schultz's contributions to the literature on economic development.

Schultz is also known for his unwillingness to bow to political pressure within agriculture during a controversy over the wartime promotion of margarine, in place of butter, in a pamphlet by an Iowa State agriculture economist, Oswald H. Brownlee. Schultz, then department chair at Iowa State, resigned in protest and took a position at the University of Chicago when Brownlee was forced to retract his pamphlet. The incident also illustrates the role of agricultural economists in providing society with objective analyses of the benefits and costs of alternative programs, policies, and investments.

D. Gale Johnson (1947), a colleague of Ted Schultz at the University of Chicago, made significant contributions in the analysis of commodity price policy, including his seminal work *Forward Prices for Agriculture*. He also contributed work on the agricultural labor market and to theoretical and practical approaches to understanding agricultural supply, among other topics. In addition to his original research, he is renowned for his contribution to the field in terms of educating students. Collected papers of Johnson can be found in *The Economics of Agriculture: Volume 1: Selected Papers of D. Gale Johnson* (Antle & Sumner, 1996a). *The Economics of Agriculture: Volume 2: Papers in Honor of D. Gale Johnson* (Antle & Sumner, 1996b) contains tributes from his former students.

Heady (1952), a professor at Iowa State University, made enormous contributions to agricultural production economics and agriculture finance. His seminal work, *Economics of Agricultural Production and Resource Use,* laid the foundation for production economics efforts for the decades following its publication. He pioneered the use of operations research techniques in farm management and made significant contributions to analysis of risk and uncertainty in agriculture.

Cochrane (1958), a professor and dean at the University of Minnesota, made substantial contributions in the field of agricultural policy and prices. His book *Farm Prices: Myth and Reality* laid out his famous treadmill theory, under which rapid output-enhancing technological advances caused a long period of production disequilibrium with resulting low product prices in agriculture. In addition to his scholarly work, he served as an agricultural advisor during John Kennedy's 1960 presidential campaign and subsequently as the chief economics advisor to the Secretary of Agriculture.

John Kenneth Galbraith (1958, 1995), a two-time winner of the Presidential Medal of Freedom, is known both for his work as a political advisor in several administrations and for his best-selling works, including *The Affluent Society*. He also published a volume laying out his largely Keynesian monetary views, *Money: Whence It Came, Where It Went.*

## Future of the Field

If the past is a predictor of the future, very likely there will be fewer agricultural economists working on a wider array of problems. Increasingly, departments and degree programs have adopted the name *applied economics* or *resource economics* or other designations, showing the shift away from a sole focus on agricultural issues. As Perry (2010) pointed out, most of these departments have not merged into the general economics department, and he predicts that the number of what he called *AE cluster* departments will not change greatly in the future but that even fewer of these departments will be known as agricultural economics.

Regardless of what their home department is called, agricultural economists will very likely continue to lend their talents to a variety of applied problems. Sandra Batie (2008) in her fellows address for the AAEA discussed "wicked problems": complex problems that do not lend themselves to the type of rational, linear approach typical of applied science. To remain relevant, she suggests that agricultural economists will need to work on these sorts of issues, stepping outside disciplinary bounds to do so. Agricultural economics, which by its nature is interdisciplinary, may be a field well placed to address "wicked problems."

In their principles of economics textbook, Paul Samuelson and William Nordhaus (2005), in a section titled "Cool Heads at the Service of Warm Hearts," stated, "The ultimate goal of economic science is to improve the living conditions of people in their everyday lives" (p. 6). As a field, agricultural economics has long embodied this view, whether the work falls under the traditional area of farm management or in one of the newer concentrations, such as health economics. A solid training in economic theory coupled with sound quantitative techniques will serve the profession well in the decades to come.

*Author's Note:* The author gratefully acknowledges the helpful comments of Henry Kinnucan, James Novak, Rachel Smith, and Michael Wetzstein on an earlier version of this chapter.

## References and Further Readings

Agrawal, R. C., & Heady, E. O. (1972). *Operations research methods for agricultural decisions.* Ames: Iowa State University Press.

Agricultural and Applied Economics Association. (n.d.). *AAEA online.* Available at http://www.aaea.org

Anderson, J. R. (1974). Simulation: Methodology and application in agricultural economics. *Review of Marketing and Agricultural Economics, 42,* 3–55.

Antle, J. M., & Sumner, D. A. (Eds.). (1996a). *The economics of agriculture: Vol. 1. Selected papers of D. Gale Johnson.* Chicago: University of Chicago Press.

Antle, J. M., & Sumner, D. A. (Eds.). (1996b). *The economics of agriculture: Vol. 2. Papers in honor of D. Gale Johnson.* Chicago: University of Chicago Press.

Barry, P. J., Ellinger, P. N., Hopkin, J. A., & Baker, C. B. (2000). *Financial management in agriculture* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Batie, S. S. (2008). Wicked problems and applied economics. *American Journal of Agricultural Economics, 90,* 1176–1191.

Beattie, B. R., & Taylor, C. R. (1985). *The economics of production.* New York: John Wiley.

Bishop, C. E., & Hoover, D. M. (n.d.). *ARE history.* Retrieved December 30, 2008, from http://www.ag-econ.ncsu.edu/ARE_history.htm

Carpio, C. E., Wohlgenant, M. K., & Safley, C. D. (2008). A structural econometric model of joint consumption of goods and recreational time: An application to pick-your-own fruit. *American Journal of Agricultural Economics, 90,* 644–657.

Chambers, R. G. (1988). *Applied production analysis: A dual approach.* New York: Cambridge University Press.

Chavas, J.-P., & Holt, M. T. (1990). Acreage decisions under risk: The case of corn and soybeans. *American Journal of Agricultural Economics, 72,* 529–538.

Chite, R. M. (2008, January). *Farm bill budget and costs: 2002 vs. 2007. CRS report for Congress.* Retrieved January 5, 2009, from http://www.nationalaglawcenter.org/assets/crs/RS22694.pdf

Cochrane, W. W. (1958). *Farm prices: Myth and reality.* Minneapolis: University of Minnesota Press.

Cochrane, W. W. (1983). *Agricultural economics at the University of Minnesota 1886–1979.* St. Paul: University of Minnesota, Department of Agricultural and Applied Economics.

College Board. (n.d.). [College search]. Available at http://www.collegeboard.com

Deaton, A. S., & Muellbauer, J. (1980). An almost ideal demand system. *American Economic Review, 70,* 312–326.

Edmondson, W. (2004). Economics of the food and fiber system. *Amber Waves, 2*(1), 12–13.

Edwards, C. (1959). Resource fixity and farm organization. *Journal of Farm Economics, 41,* 747–759.

Eisgruber, L. M., & Reisch, E. (1961). A note on the application of linear programming by agricultural economics departments of land grant colleges. *Journal of Farm Economics, 43,* 303–307.

Erickson, S. P., Akridge, J. T., Barnard, A. F., & Downey, W. D. (2002). *Agribusiness management* (3rd ed.). New York: McGraw-Hill.

Forker, O. D., & Ward, R. W. (1993). *Commodity advertising: The economics and measurement of generic programs.* New York: Lexington Books.

Forward. (1919). *Journal of Farm Economics, 1,* 1–2.

Galbraith, J. K. (1958). *The affluent society.* Boston: Houghton Mifflin.

Galbraith, J. K. (1995). *Money: Whence it came, where it went.* Boston: Houghton Mifflin.

Gardner, B. L., & Rausser, G. C. (Eds.). (2001a). *Handbook of agricultural economics: Vol. 1A. Agricultural production.* New York: Elsevier North-Holland.

Gardner, B. L., & Rausser, G. C. (Eds.). (2001b). *Handbook of agricultural economics: Vol. 1B. Marketing, distribution and consumers.* New York: Elsevier North-Holland.

Goodwin, J. W. (1994). *Agricultural price analysis and forecasting.* New York: John Wiley.

Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technical change. *Econometrica, 25,* 501–522.

Hardaker, J. B., Huirne, R. B. M., Anderson, J. R., & Lien, G. (2004). *Coping with risk in agriculture* (2nd ed.). Cambridge, MA: CABI.

Heady, E. O. (1946). Production functions from a random sample of farms. *Journal of Farm Economics, 28,* 989–1004.

Heady, E. O. (1952). *Economics of agricultural production and resource use.* Englewood Cliffs, NJ: Prentice Hall.

Heady, E. O., & Candler, W. (1958). *Linear programming methods.* Ames: Iowa State University Press.

Houck, J. P. (1964). A statistical model of the demand for soybeans. *American Journal of Agricultural Economics, 46,* 366–374.

Houck, J. P., & Ryan, M. E. (1972). Supply analysis for corn in the United States: Impact of changing government programs. *American Journal of Agricultural Economics, 54,* 184–191.

Huang, S., & Sexton, R. J. (1996). Measuring returns to an innovation in an imperfectly competitive market: Application to mechanical harvesting of processing tomatoes in Taiwan. *American Journal of Agricultural Economics, 78*(3), 558–571.

International Association of Agricultural Economists. (2008). [Home page]. Available at http://www.iaae-agecon.org

Johnson, D. G. (1947). *Forward prices for agriculture.* Chicago: University of Chicago Press.

Just, R. E. (2008). Distinguishing preferences from perceptions for meaningful policy analysis. *American Journal of Agricultural Economics, 90,* 1165–1175.

Just, R. E., & Pope, R. D. (Eds.). (2002). *A comprehensive assessment of the role of risk in U.S. agriculture.* Norwell, MA: Kluwer Academic.

Kay, R., Edwards, W., & Duffy, P. (2008). *Farm management* (6th ed.). New York: McGraw-Hill.

Knutson, R. D., Penn, J. B., & Flinchbaugh, B. L. (2007). *Agricultural and food policy* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Kohls, R., Uhl, J., & Hurt, C. (2007). *Marketing of agricultural products* (10th ed.). Upper Saddle River, NJ: Prentice Hall.

LaFrance, J. T., & Pope, R. D. (2008). Homogeneity and supply. *American Journal of Agricultural Economics, 90,* 606–612.

Nerlove, M. (1956). Estimates of the elasticities of supply of selected agricultural commodities. *Journal of Farm Economics, 38,* 496–509.

Nichols, J. P., Kinnucan, H. W., & Ackerman, K. Z. (Eds.). (1990). *Economic effects of generic promotion programs for agricultural exports.* College Station: Texas A&M University.

O'Donnell, C. J., Shumway, R. C., & Ball, V. E. (1999). Input demands and inefficiency in U.S. agriculture. *American Journal of Agricultural Economics, 81,* 865–880.

Penson, J. B., Jr., & Hughes, D. W. (1979). Incorporation of general economic outcomes in econometric projections models for agriculture. *American Journal of Agricultural Economics, 61,* 151–157.

Penson, J. B., Capps, O., Jr., Rosson, C. P., & Woodward, R. (2006). *Introduction to agricultural economics* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Perry, G. (2010). What does the future hold for departments of agricultural economics? *Applied Economic Perspectives and Policy, 32*(1), 117–134.

Piggott, R. (1992). Some old truths revisited. *Australian Journal of Agricultural Economics, 36,* 117–140.

Samuelson, P. A., & Nordhaus, W. D. (2005). *Economics* (18th ed.). New York: McGraw-Hill.

Schultz, T. W. (1946). Changes in economic structure affecting American agriculture. *Journal of Farm Economics, 28,* 15–27.

Schultz, T. W. (1964). *Transforming traditional agriculture.* New Haven, CT: Yale University Press.

Shaars, M. A. (1972). *The story of the Department of Agricultural Economics 1909–1972.* Retrieved December 10, 2008, from http://www.aae.wisc.edu/pubs/AAEStory.pdf

Shumway, R. C. (1983). Supply, demand, and technology in a multiproduct industry: Texas field crops. *American Journal of Agricultural Economics, 65,* 748–760.

Spillman, W. J. (1923). Application of the law of diminishing returns to some fertilizer and feed data. *Journal of Farm Economics, 5,* 36–52.

Stanton, B. F. (2001). *Agricultural economics at Cornell: A history, 1900–1990.* Ithaca, NY: Cornell University, College of Agriculture and Life Sciences.

Stigler, G. W. (1945). The cost of subsistence. *Journal of Farm Economics, 27,* 303–314.

Tauer, L. (1983). Target MOTAD. *American Journal of Agricultural Economics, 65,* 606–610.

Tintner, G., & Brownlee, O. H. (1944). Production functions derived from farm records. *Journal of Farm Economics, 26,* 566–571.

Tomek, W. G., & Robinson, K. L. (2003). *Agricultural product prices* (4th ed.). Ithaca, NY: Cornell University Press.

Warren, G. F. (1913). *Farm management.* New York: Macmillan.

Waugh, F. V. (1928). Quality factors influencing vegetable prices. *American Journal of Agricultural Economics, 10,* 185–196.

Wohlgenant, M. K. (1989). Demand for farm output in a complete system of demand functions. *American Journal of Agricultural Economics, 71,* 241–252.

Yeager, J. C., & Stevenson, G. (2000). *Inside Ag Hill: The people and events that shaped Auburn's agricultural history from 1872 through 1999.* Auburn, AL: Auburn University, College of Agriculture.

Zusman, P., & Amiad, A. (1965). Simulation: A tool for farm planning under conditions of weather uncertainty. *Journal of Farm Economics, 47,* 574–594.

# 59

# Real Estate Economics

Erick Eschker
*Humboldt State University*

Real estate economics is the study of the markets for land and structures, including residential housing, commercial office space, and industrial warehouses. Although much theory has existed for decades regarding real estate markets, questions are being answered today regarding mortgage finance innovation, the rise of suburban business and residential centers, and the effects of zoning laws. The housing boom and bust of the first decade of the twenty-first century has meant that households, businesses, and governments have become more interested than ever in real estate and its effect on various aspects of the economy. In this chapter, the theory of real estate markets is first presented, including the relevance of real estate rental price, the heterogeneous nature of real estate, the importance of location, and the interaction with the macroeconomy. Next, various contemporary applications of real estate economics are discussed, followed by a description of government policy intervention in the real estate market. Finally, the housing boom and bust is presented along with areas for future research.

## Theory of Real Estate Markets

### Flow of Services, Rent, and Price

In many respects, real estate markets are similar to markets for other goods and services. There exist buyers who demand real estate by demonstrating a willingness and ability to pay for property. There are also sellers who supply real estate. One unique feature of real estate is that the goods in question—namely, land and structures—are long-lived. Consuming real estate does not result in the disappearance of the good as with, say, consuming a slice of pizza. Real estate can be purchased and enjoyed today and then sold again tomorrow. In fact, the vast majority of real estate sold in any year was previously owned. Because of this durability, the decision to buy or sell real estate must take a long time horizon into account.

Durable goods deliver a flow of services over time to the owner or user of that good. For example, a car lasts many years, and a car will deliver a flow of transportation services each year. Real estate delivers a flow of shelter services, in residential housing, and a flow of retail store space, in commercial real estate. The person who purchases a durable good takes current and all future flows of services into account when making the decision to purchase the good. This is true whether or not the buyer intends to use the services that flow from the durable good. For example, in the case of a car, although often the buyer intends to drive the car for a number of years after the purchase, car rental companies purchase large amounts of cars in order to charge rent to their customers. The same is true in real estate, where a large amount is owner occupied, while other people intend primarily to offer for rent the shelter or retail space to others. If the buyer intends to rent the real estate to others, then the buyer's willingness to pay will be based on the rent that can be earned, which is determined by renters and landlords in the rental market for real estate. Renters will be willing to pay rent that is at or below the value that the renter puts on the flow of services. In fact, the dollar value of rent is an easy-to-obtain measure of the dollar value of the flow of services from a property.

Purchasing a piece of real estate today gives the buyer the use of services this year and every year in the future. See James Hamilton (2005) for a good discussion about how the buyer may approach the decision of whether to buy a house. The willingness to pay for that stream of service

flows is equal to its present discounted value. The present discounted value is the sum of the discounted values, where the appropriate discount rate reflects the buyer's opportunity cost of purchasing today. Typically, it is the mortgage interest rate, which may be adjusted for risk and taxes. Numerically, the present discounted value in year $t$, $PDV_t$, is as follows:

$$PDV_t = S_t + \frac{S_{t+1}}{(1+i)} + \frac{S_{t+2}}{(1+i)^2} + \cdots, \qquad (1)$$

where $S_t$ is the flow of real estate services, $i$ is the interest rate, and $t$ is the year index. If the value of services grows at a constant rate $g$, and if $g < i$, then the above reduces to

$$PDV_t = \frac{S_t(1+i)}{(i-g)}. \qquad (2)$$

The present discounted value will rise with the value of the flow of services provided by the property and the growth rate of services, and it will fall when the interest rate rises. This last result is because an increase in the interest rate will not only raise the numerator but also raise the denominator more, thus lowering the ratio. The present discounted value of the flow of current and future services is referred to as the *fundamental value* of real estate. The fundamental value simply says that the most that people will pay for a piece of real estate is the present discounted value of the flow of services. This is a very intuitive way to think about real estate pricing and shows the link between the market for real estate rentals and the market for real estate purchases—the main driver of real estate prices is the value that the end user puts on the property and the discount rate. The fundamental value of real estate is essentially an arbitrage condition that says that if the market price of real estate is equal to the present discounted value of future rents, then there are no sure profits to be made through buying or selling real estate, and thus, the market price of real estate will not rise or fall unless there is a change to the fundamental value.

## Demand and Supply

The price of real estate in the short run is determined by supply and demand in the real estate market. See Denise DiPasquale and William Wheaton (1996, chap. 1) for a rigorous exposition of this material. The market for real estate, where potential buyers and sellers meet, is different from the market for rental properties, where renters and landlords come together to rent buildings. This discussion focuses on the market for rental properties and considers how that market sets the price and influences the construction, or development, of new real estate. In the rental market, each potential renter has an individual willingness to pay. The market willingness to pay is the sum of the willingness to pay for all renters, and just as with any good, the sum of willingness to pay determines the market demand for the good. As the market rental price of real estate falls, more and more renters will rent, and this results in the typical downward sloping demand curve, as shown by line $\text{Demand}_1$ in Figure 59.1. Market demand is a function of the size and composition of the population, the local job market and incomes, local amenities, and other factors that are discussed later. Demand shifts come about when any of these factors change. The quantity or stock of real estate is fixed in the short run because it takes a relatively long time for new real estate to be built or developed, given permitting, zoning, and construction considerations. Figure 59.1 therefore shows supply, $\text{Supply}_1$, in the rental market as a vertical line equal to the stock of real estate.

Because real estate is long-lived, today's supply is based on how much property was developed in the past. The amount of real estate construction or investment in any year

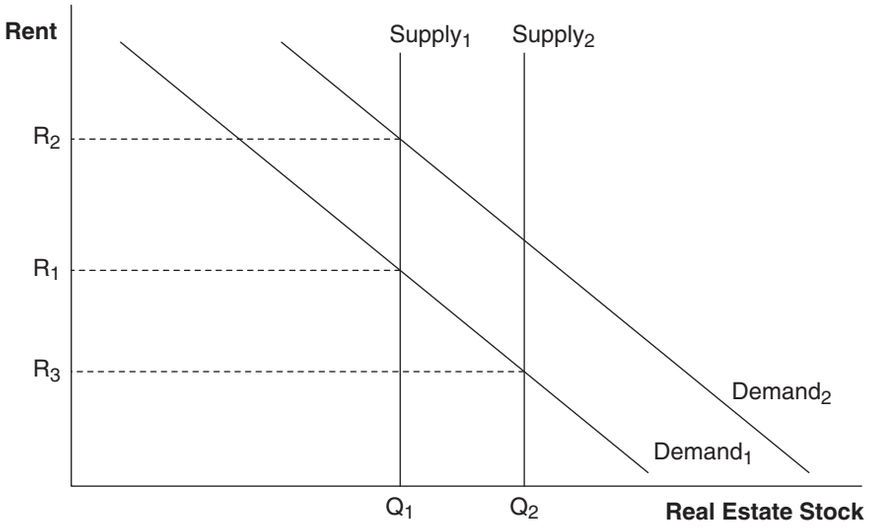


**Figure 59.1** Real Estate Rental Market

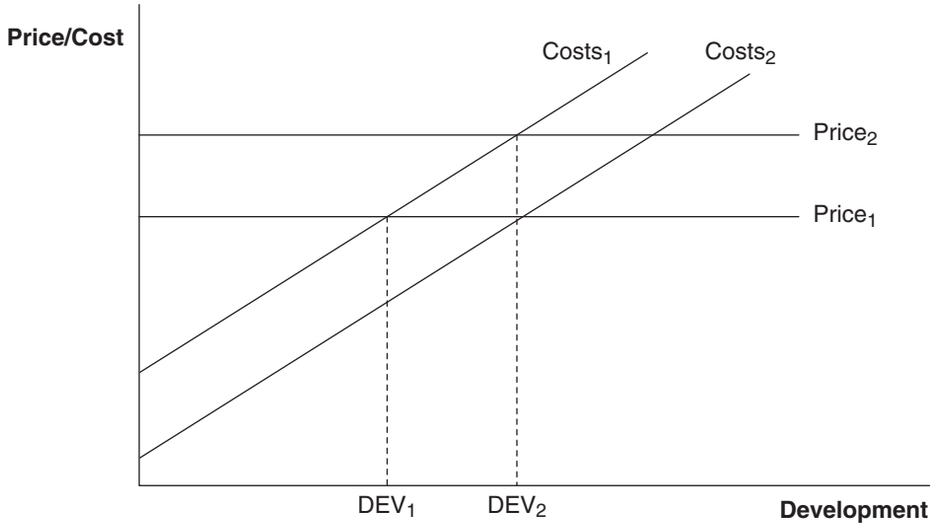


**Figure 59.2** Real Estate Development

depends on the availability and cost of funds for construction, on construction costs and technology, and on zoning considerations. Construction costs include fixed and marginal costs, and an increase in these costs will lead to less investment. Investment also depends on the price of real estate at the time that the investment decision is made. All else the same, a higher price of real estate will induce more construction and will lead to a higher future stock of real estate. If the price of real estate is above construction costs, then developers will find it profitable to build, but if the current price of real estate is below construction costs, then property will not be developed. Figure 59.2 shows how real estate development is determined. Given the price of real estate, there is one level of development that will be obtained. In Figure 59.2, if the price is equal to $Price_1$, and if costs are represented by line $Costs_1$, then real estate development is equal to $DEV_1$. If this new construction is greater than depreciation, which is the erosion of the real estate stock due to normal wear and tear, then the supply of real estate in the rental market will be greater next period. Real estate is just like any capital good in that the stock of real estate grows only if the investment flow is greater than depreciation.

There is one rental price at which the quantity of real estate supplied is equal to the quantity demanded, and that is the market clearing or equilibrium rent. In Figure 59.1, if the initial demand is represented by the line $Demand_1$, and if the initial supply is represented by the line $Supply_1$, then the initial equilibrium rent is $R_1$. This current rent, along with the expected growth of rent and mortgage interest rates, will determine the present discounted value of future rents and the price that people are willing to pay to buy the real estate, and will it determine real estate development.

The equilibrium rent will change when demand or supply shifts. For example, suppose that population grows in a particular market. This will shift to the right the demand for real estate to $Demand_2$, pushing up the rent to $R_2$ in Figure 59.1. This increase in rent will lead to an increase in price to $Price_2$ in Figure 59.2 and will lead to increased construction level $DEV_2$. In the end, rent is higher, the price level is higher, and there is a larger stock of real estate. As another example, suppose that demand and price are back to their original positions and that credit becomes more easily available to developers. This will reduce construction costs to $Costs_2$ in Figure 59.2 and increase investment to $DEV_2$. The supply of real estate in Figure 59.1 will rise to $Supply_2$, which lowers the rent to $R_3$ and eventually lowers the price of real estate. In the end, the rent is lower, the price is lower, and there is a larger stock of real estate.

It is worth mentioning that although changes in the market price and the rental price of real estate are often positively correlated, as shown in the preceding two examples, one exception is when the mortgage rate changes. If the mortgage interest rate rises, this will decrease the present discounted value of future rents and lower the price. As a result, there will be less development, and the supply of real estate will fall in the rental market, which leads to higher rents.

## Heterogeneity

Real estate is not a homogenous good in which each parcel of real estate is indistinguishable from another. Instead, each parcel has unique characteristics, including location, size, and amenities, and demand for these unique characteristics is determined in the market. The value for a particular real estate parcel is based on its unique characteristics. For example, people are generally willing to pay more to live near the ocean, and houses that are near the ocean have an ocean premium built into their market price. The more desirable features that a parcel has, the higher

will be its market price. However, despite the uniqueness of each parcel, demand for a particular parcel is usually price elastic because other similar parcels are substitutes.

The demand and supply for a particular piece of real estate is composed of two distinct types of factors. The first is those characteristics that are unique to that parcel, such as exact location, number of rooms, size, land use zoning, and age of the structure. These unique characteristics will lead to different market prices among pieces of real estate. For example, commercial real estate that has better access to major road arteries will command a higher market rent and thus a higher market price, all else the same. But there are also factors that are common to all real estate within a certain market. In fact, one can define the market as all the real estate affected by these common factors. These common factors include the population, the local unemployment rate, and proximity to outdoor recreation activities, among others. These common factors will tend to raise or lower the market price of all properties but not change the relative prices. For example, if a large employer leaves the area, then this will decrease rents and lower the demand and prices for all real estate within the market.

## Real Estate Value and Location

The theory of real estate value begins with the classic concept of Ricardian rent, which predicts that more desirable real estate will command a higher rent and market price as users compete for land. In the monocentric city model, distance to the single center of the city largely explains real estate prices. In urban residential real estate pricing, workers in a city must commute to their places of work, and they are willing to pay more for housing that is closer to their employers to avoid costs of commuting. Thus, land at the city center has a higher market price than land at the city edge. The price of all housing, which is made up of the price of the land and the price of the house, must be high enough to bid resources away from agricultural and physical structure uses. Thus, at the city edge, house prices must equal the sum of returns to agricultural land and construction costs of the house structure. But housing closer to the city center will have a higher price because people will pay to avoid the commute costs associated with greater distance from the city center. As commuting costs rise, prices in the city center rise relative to prices at the city edge. Users of urban land who differ based on their commuting costs, such as those with high wages and thus high opportunity costs of commuting time, will segregate themselves by location. Those with higher commuting costs will live closer to the city center, and if they earn higher wages, then segregation will appear along income levels. Because property closer to the city center has higher rent gradients associated with it, builders will often increase density to substitute structure for land, and residential real estate closer to the city center will be more densely populated. The original city center was often built decades or centuries earlier and often has very low density. Redevelopment of the city center into higher density use will become profitable for developers once residents feel that reduced commute costs are large enough to outweigh the dislike of higher density living.

This monocentric model of urban development has many shortcomings, as outlined in Alex Anas, Richard Arnott, and Kenneth Small (1998). The most visible is the fact that many of today's cities are best described as multicentered, with suburban city centers surrounding the original city core. Industrial land often occupies land on the edge of the city where land rent is cheaper and truck access to roads is good. Both industrial and commercial land has gathered in new outlying clusters within larger metropolitan regions. These alternative city centers compete with the traditional city hub for workers, and there has been much employment growth in these alternative suburban centers. A strong incentive to pull firms outside the central city hub is lower wages in the suburbs, where more and more of its workforce lives and where commuting times and thus wages are lower. Economies of agglomeration refer to the reduced costs to firms by clustering near one another. These include access to a large pool of skilled workers, better communication between firms, and closer proximity to suppliers. Counteracting this migration to the suburbs are increased costs of isolation and the reduction in agglomeration economies. At some point, the draw to an alternative city center is greater than the loss of the agglomeration effects, and the firm leaves the city center. One expects firms with the largest economies of agglomeration to locate in the large city center, while firms with less economies of agglomeration will be found in the alternative city centers. Recent increases in information technology, such as the Internet, may reduce the benefits to firms of clustering near one another, but Jess Gaspar and Edward Glaeser (1998) suggest that telecommunications may be a complement for face-to-face interaction and promote city center growth.

Commercial real estate can profit by locating near customers, suppliers, roadways, or mass transit terminals. Retailers would also like to cluster near complementary retailers because consumers try to minimize travel costs and the number of trips and often prefer to shop at a single shopping center. However, retailers would like to be located farther from similar retailers, because that increases their market power and pricing ability. Shopping malls give retailers the chance to cluster together in a central location, which may benefit some retailers, although some retailers, such as those who sell lower priced, nondurable goods that are purchased frequently are often not found in central shopping malls. Thus, rent for retail and commercial space is based partly on the mix of businesses in the area and the likelihood that those businesses will draw customers to the firm. See Eric Gould, B. Peter Pashigian, and Canice Prendergast (2005) for a recent discussion of shopping mall pricing.

### The Macroeconomy and Real Estate

The real estate market is both a determinant of the regional macroeconomic climate and a product of that climate. Land and buildings are inputs into the production process, whether those markets are for industrial, commercial, or residential purposes, and real estate prices affect costs of production. Exogenous increases in the supply of real estate, perhaps from the opening of new lands for development, reduce rents and reduce costs to firms in the region, which gives those firms a cost advantage over other regions of the state or country. Employment and wages in the region will increase. A very inelastic supply of real estate can inhibit demand-driven economic growth if increases in housing prices raise the cost of living sufficiently to keep real wages from rising and attracting new workers into a region. Economy-wide factors, such as an exogenous increase in labor supply due to immigration, will affect the real estate market. Immigration will increase employment and production because wages and costs to firms fall, and it will increase the demand for real estate, which raises real estate rents. An increase in the demand for goods produced in the region will lead to an increase in the demand and price of all factors of production, including real estate. If the supply of real estate is price elastic, real estate prices will increase less and real estate development will be more than if real estate supply is inelastic. Local real estate markets depend on the national economy, with recessions typically reducing the demand for most types of real estate. But regional differences matter too, and a local real estate market may suffer relatively less during a recession if the relative demand for the particular mix of goods and services produced in the region does not fall. Finally, real estate wealth is a large fraction of total household wealth, and Karl Case, Robert Shiller, and John Quigley (2005) find large impacts of housing wealth on consumption.

## Applications and Empirical Evidence

### Hedonic Prices

Richard Muth (1960) wrote an early paper on hedonic pricing. A hedonic price equation relates the price of a piece of real estate to various characteristics of the property, such as distance to city center, size measured in square feet, number of bathrooms, and neighborhood quality, which can be measured as either low quality (0) or high quality (1). This is an attempt to resolve the issue that real estate is not a homogeneous good. The hedonic price equation decomposes real estate into various characteristics and estimates prices for each characteristic. In the case of residential housing, an example is as follows:

$$Price = \alpha + \beta_1 DISTANCE + \beta_2 SIZE + \beta_3 ROOMS + \beta_4 NEIGHBORHOOD, \quad (3)$$

where $\alpha$ is a constant and the $\beta$ s are the coefficients on each housing characteristic. Each $\beta$ is the marginal impact of increasing the characteristic by one unit, in continuous variables such as distance, or the marginal impact of having that characteristic present, in binary characteristics such as neighborhood quality. One would expect that the coefficient on distance to the center of the city to be negative, because the amount of rent that people are willing to pay falls the farther the house is from the center of city, all else the same. One should note, however, that in very large metropolitan areas, some suburbs have developed into their own city centers, and distance to these new city centers may be just as important to house price as distance to the historic city center. James Frew and Beth Wilson (2002) present an empirical estimation of rent gradients in the multicentered city of Portland, Oregon. The coefficients on size, number of bathrooms, and neighborhood quality are expected to be positive because people are willing to pay greater rent for housing as these features increase, and thus, market real estate prices will be higher.

The coefficients in hedonic price equations are frequently estimated for both residential and commercial and industrial real estate parcels using ordinary least squares regression analysis or another estimation technique. There are typically many characteristics used to explain the price of real estate. In residential housing, these include local school quality, nearness to employment centers, proximity to amenities such as the ocean or recreational facilities, yard size, presence of a garage, and age of the house structure. In commercial or industrial property, these characteristics include nearness to major highways, airports, and rail lines; amount of pedestrian or car traffic; zoning considerations; nearness to suppliers, distributors, and customers; and parking availability, among others. Ioan Voicu and Vicki Been (2008) look at the effect of community gardens on property values in New York City and find a significant effect for high-quality gardens.

### Demographics and Housing

The demand for residential real estate depends in the long run on characteristics of the population that move slowly over time. Gregory Mankiw and David Weil (1989) present a look at the implications of the baby boomers and the aging of the population on housing supply and demand. One factor is net household formation, which is the difference between newly created households and newly dissolved households in a year. A household, which is a group of related or unrelated individuals living at the same parcel of real estate, is a common unit of measurement by the Census Bureau. New households may be created when children leave their parents' residence and through divorce, for example, while the number of households may fall during marriage and death, for example. The average size of households, as well as the age, composition, and income of households, will influence the typical features found in

newly constructed housing, because developers quickly respond to current market conditions.

Age, income, demographic mobility, and marital status also greatly influence whether the household rents or buys housing. According to the 2007 American Housing Survey (U.S. Census Bureau, 2008), 31% of households rent the houses that they live in, 24% own the houses outright, and 44% are making mortgage payments. Renters are on average younger and have lower incomes than nonrenters. Permanent, or long-run average income, appears to influence the decision on whether to buy more than current income. Transaction costs of buying and selling a house are large, and renters move much more frequently than owners. Getting married and having children appear to be big factors that explain the switch from renting to owning.

Housing vacancies are a closely watched measure in the housing market. For a given level of sales, as vacancies rise, the average time on the market increases. Increased time on the market will tend to lower market prices, because sellers face increased opportunity cost of funds if they cannot sell their houses. New house builders will also be more motivated to lower prices, and they will also respond to greater average time on the market by reducing new construction. Although vacancy rates are typically studied at the local market level, consider the recent drop in U.S. new housing demand. The U.S. Census Bureau (2009) reports that the U.S. homeowner vacancy rate increased very quickly over 2006, averaging 2.375 in 2006 compared to 1.875 in 2005. New house sales fell from 1.283 million in 2005 to 1.051 million in 2006. In December of 2005, there were 515,000 new houses for sale, and the median time for sale was 4.0 months, while in December of 2006, there were 568,000 new houses for sale, and the median time for sale was 4.3 months. The median price of new houses continued to rise for 2 more years, until falling from $247,900 in 2007 to $230,600 in 2008. Builders responded to these market changes, and from 2005 to 2006, housing starts fell from 2.068 million to 1.801 million units.

### Mortgage Financing

Residential real estate typically sells for hundreds of thousands or even millions of dollars. This is greater than most households' annual incomes. Households typically do not save over many years to purchase housing but rather obtain financing to purchase residential real estate. Financing terms have changed considerably over the last 10 years and continue to change. Except for a short number of years in the first decade of the 21st century when houses could be bought with little or no down payment, down payments of 20% are common on loans. The down payment is subtracted from the sum of the purchase price and all fees and expenses to determine how much the buyer must finance. A mortgage is a residential real estate loan with the real estate itself as collateral. Most mortgages are fully amortizing, which means that the principal balance is repaid over the life of the loan, which is usually 30 years but sometimes more or fewer, such as 40 or 15 years. A common question asked is why does the principal balance decline very slowly in the first years of the loan? To answer this, it is helpful to think of a mortgage as a loan that is to be repaid every month. After the first month of the loan, the interest due is extremely large, because the outstanding principal is large. As the number of months goes by, the interest due becomes less because the principal is being paid down. In the final months, the interest due is small because the principal is small.

Many residential mortgages do not have fixed mortgage interest rates over the lives of the loans but rather have a rate that rises and falls, within limits, along with the prevailing interest rate. These adjustable rate mortgages (ARMs) became more popular with lenders after the high inflation of the 1970s, when lenders were paying high interest rates on short-run liabilities (deposits) while receiving much lower interest rates on long-term assets (loans). Another recent change to mortgages has been the use of mortgages that do not fully amortize. These interest-only and even negative-amortizing loans allow the monthly mortgage payments to vary, within limits, such that the loan principal balances do not fall as they would with fully amortized loans. These pick-a-payment mortgages usually require full amortization to begin a few years after the loan begins.

Funding for mortgages is typically provided by commercial banks and thrifts, and these institutions, as well as mortgage brokers, typically issue mortgage loans to households. However, it is very common for these loans to be quickly sold on the secondary mortgage market to pension funds, insurance companies, and other investors. Mortgage-backed securities (MBSs) or mortgage bonds are sold to investors. Investment banks and the government-sponsored enterprises Fannie Mae and Freddie Mac package individual mortgages into these securities, which are sold on huge markets. The pooling of individual mortgages can reduce risk because only a small number of borrowers are expected to default. Fannie Mae and Freddie Mac also guarantee the payments on the securities that they process. Mortgages that conform to Fannie Mae and Freddie Mac standards, such as maximum loan to value, make up most of the secondary market, and the standardization of loans has allowed the market for these securities to grow. Conforming mortgage loan amounts must be below a threshold limit, adjusted to take average market price into account, or else the loans may be classified as jumbo loans, which have higher interest rates.

## Government Policy

### Tax Treatment of Housing

The federal government provides incentives to purchase housing to encourage home ownership. Government intervention is often justified on the grounds of positive externalities that result from home ownership, such as the idea

that home owners will better maintain the exterior appearance of their houses than home renters or landlords. Glaeser and Jesse Shapiro (2003) lead an excellent discussion about the positive and negative externalities associated with home ownership. The mortgage interest deduction, whereby interest payments (not principal payments) are tax deductible, cost the federal government $67 billion in 2008, according to the Joint Committee on Taxation (2008). Additionally, local property taxes are deductible for federal income taxes. The marginal benefit to the owner of these tax deductions is equal to the deduction multiplied by the marginal tax rate. Because federal taxes allow a standard deduction, the housing tax incentives are relevant only if the household itemizes, but low- to moderate-income households often do not itemize. The evidence shows that federal tax treatment of housing has a more powerful effect in encouraging people to purchase larger houses rather than to purchase housing for the first time. Other tax incentives include a large exemption on housing capital gains and the mortgage interest deduction for state income taxes in many states. According to the Legislative Analyst's Office (2007) in California, the largest California income tax deduction is the mortgage interest deduction, which cost $5 billion in reduced revenue in 2007 to 2008. Various distortions are created by state tax treatment of housing, such as Proposition 13 in California, which discourages sales of existing houses because property taxes tend to be lower the longer one resides at an address.

## Retail Development

In many towns, the location of newly developed retail space is a very sensitive political matter that often requires a vote by the city council or a voter referendum. This is because of the impact of new retailers on the sales of existing retailers. If a new shopping center is built, existing retailers will often lose sales and market pricing power. Of course, the total amount of retail sales may remain unchanged, with the new shopping mall simply taking sales from existing business, and sales in one town may rise at the expense of sales in a nearby town. In rural communities, a new shopping center may divert sales away from out-of-the-area retailers or Internet retailers. The rise of Internet retailing can reduce the overall demand for retail real estate while increasing the demand for warehouse real estate. It remains to be seen how demand for retail space, and retail rental prices, will be affected in the long run by the rise in Internet retailers. Elaine Worzala and Anne McCarthy (2001) present an early examination of Internet retailing on traditional retail location.

## Zoning

Municipalities, counties, and special local governments, such as water districts, Indian tribes, and coastal commissions, set property tax rates and zone land for specific uses. They use this tax revenue and grants from the federal and state governments to administer public expenditures on items such as schools and infrastructure. These local governments usually have a much larger impact on real estate markets than do the federal and state governments. In the classic Tiebout model, the difference in spending levels by communities on public services, such as school quality and police protection, reflects differences in community preferences. Communities that value higher quality schools will spend more on school services than those communities that do not, and residents will segregate into locations that reflect differences in willingness to pay for public services. Because willingness to pay often depends on income, differences in public expenditures will also reflect differences in average income across communities.

When considering new land development for either residential or commercial purposes, local governments will ask whether the new development will make existing residents better or worse off. If the new development requires on average less expenditure and raises more in property taxes compared to existing developed land, then current residents will be better off. On the other hand, if the new development requires more expenditure and raises less in property taxes on average compared to existing development, then current residents will not be in favor of development. Property taxes are based on land value and are not directly proportionate to the number of households on a parcel. It is therefore likely that compared to existing housing, denser housing development will generate a demand for public services that is greater than the higher property tax revenue collected. As a result, many localities require a minimum lot size, which has the effect of keeping communities segregated by lot size and income.

Commercial and industrial real estate is also zoned by local governments. The big concern with nonresidential development is the impact on the character of the community. Households typically want to live near other households or amenities such as parks and prefer to not live near commercial or industrial sites. This is one reason property taxes tend to be set higher for nonresidential uses. Towns that are more willing to accept new tax revenue in exchange for an altered community character are more likely to have nonresidential real estate developed within their jurisdictions.

Zoning authority is typically justified by the externalities that land use imposes on other individuals in the market and by the need to provide public goods. Of course, zoning is a much-debated political matter, and there is no doubt that in practice, zoning considerations are often made that redistribute resources rather than improve the quality of living for citizens. Zoning will restrict the supply of new development and raise real estate prices. Glaeser and Joseph Gyourko (2003) find that in U.S. regions where zoning is more restrictive, such as coastal states, high housing prices are the result. Negative externalities resulting, for example, from industrial pollution, noise, odor, dilapidation, and increased parking congestion, can adversely impact residential, commercial, and

industrial real estate values. Positive externalities such as brown field restoration or regular exterior house maintenance can increase real estate values. Transaction costs tend to make an optimal private market solution in the presence of externalities difficult to obtain, and a common economic solution is government intervention. Local governments extensively control or restrict land use to alleviate externalities. For example, a large city might require newly constructed apartments to have at least one on-site parking space per residential unit to reduce street parking congestion. Governments also turn toward market incentives that internalize the external costs, such as a large city charging car owners a fee for on-street parking permits. Another example of a serious negative externality from real estate is traffic congestion. Users of residential, commercial, and industrial real estate drive on existing roadways and can create substantial gridlock, especially at rush hour. This increases travel time for users, and estimates of the aggregate value of time lost can be substantial.

The City of London recently implemented a traffic congestion charge on cars during business hours, and the effect has been a substantial drop in congestion. Public goods, which once created are nonexcludable, face the well-known free-rider problem, which predicts underprovision of the good. Public goods include roads, parks, and footbridges. Local governments often zone real estate to provide for open space or tax residents to pay for road maintenance, for example. Richard Green (2007) finds that airports can help a region to experience economic growth. The optimum land use and government response to externalities and public goods will change over time. However, existing real estate usage is based on historical decisions, because real estate structures are typically long-lived, and governments may need to periodically revisit land use zoning.

## Real Estate After 2000 and Future Directions

### Housing Boom and Bust

Residential real estate prices began to climb very quickly in the early 2000s in many countries across the globe, and starting around 2006, prices began to decline. In the United States, nominal prices climbed 90% as measured by the Case-Shiller Index. The Case-Shiller Index is a repeat sales price index, which calculates overall changes in market prices by looking at houses that sold more than once over the entire sample period to control for differences in quality of house sold. In some regions, the median selling price of a house doubled in about 3 years from 2002 to 2005. Although regional real estate markets had previously experienced quick price increases, such as the Florida land boom of the 1920s, by all accounts this was the fastest increase in housing prices at the aggregate national level ever. As Shiller (2007) shows, inflation-adjusted U.S. housing prices were remarkably consistent since 1890, with the 1920s and 1930s being the exception of years of low prices. But since the late 1990s, housing prices in the United States on average grew to levels not seen in any year in which data are available. This fast rise in house prices encouraged a great deal of investment into both residential and commercial real estate. Construction employment soared, as did the number of people working as real estate agents and mortgage brokers. By 2008, housing prices had fallen 21% from their peak levels, which was a greater fall, in inflation-adjusted terms, than during the Great Depression. Accompanying this massive drop in prices was a fall in new and existing house sales, an increase in foreclosures to record levels, a drop in construction spending and commercial real estate investment, massive drops in government tax collections, and a huge drop in jobs in construction, real estate, and the mortgage industry. Mortgage companies, banks, and financial firms that held real estate assets shut their doors, and there was massive contraction and consolidation in these industries. The U.S. recession that began in 2007 accompanied the collapse of the housing market.

It is important to explain both the boom and the bust of this historic housing cycle. Cabray Haines and Richard Rosen (2007) compare regional U.S. housing prices and find evidence that actual prices rose above fundamental values in some markets. The boom can be partly explained by the fact that mortgage interest rates were very low at the onset, which according to the fundamental value approach should raise house prices because lower interest rates will raise the present discounted value of future rents. But it seems that changes in mortgage financing may have been an even more important part of the explanation, and many economists characterize the housing boom as a manifestation of a credit bubble that saw simple measures such as the price-to-rent and price-to-income ratios rise to levels above historic averages. Housing affordability fell sharply in most of the country. The use of MBSs increased, and mortgage lenders greatly reduced holdings of their mortgages in the so-called originate-to-distribute model. Mortgage lending standards fell as more and more people fell into the subprime (poor credit history) or Alt-A (high loan to income value) borrowing categories and down payment and income documentation requirements were greatly reduced.

Both this increase in supply and the increase in demand for mortgage credit increased the demand for houses, which pushed prices to record highs. Starting in 2006, and particularly in 2007, the appetite for MBSs fell greatly as mortgage defaults increased across the country. The drop in available mortgage credit was a blow to new buyers, those refinancing, and those seeking to trade up or purchase second houses. The result was a huge drop in demand for housing. The U.S. Census Bureau (2009) reports that in the fourth quarter of 2008, the home ownership rate was 67.5%, the same level as the fourth quarter of 2000.

The Federal Reserve's (2009) Flow of Funds Accounts report that household percentage equity was lower than at any time in over 50 years of data. Government reaction to the housing bust has been unprecedented. In 2008, the U.S. Treasury Department and the U.S. Federal Reserve Bank tried to encourage consumer confidence in financial markets, brokered mergers between insolvent institutions and other companies, and committed hundreds of billions of dollars to help the financial industry and the U.S. economy at large. In September of 2008, Fannie Mae and Freddie Mac were placed into government conservatorship when their stock prices plummeted and they were unable to raise capital. A flurry of legislation passed that reduced income taxes for home buyers. At the federal level, first-time homebuyers in 2008 who were below the income limit received a credit on their taxes up to $7,500, which must be repaid over 15 years, while in 2009, there was a refundable tax credit equal to 10% of the purchase price of the house, up to $8,000. In California, a non-means-tested tax credit of 5%, up to $10,000, was available to all purchasers of newly constructed houses. Other states had their own home buyer income tax credits.

### Areas for Future Research

One of the challenges of future research will be to better understand how financial product innovation contributed to the housing boom and what types of regulation would best fend off a similar future boom and bust cycle. Chris Mayer, Karen Pence, and Shane Sherlund (2008) point to zero-down-payment financing and lax lending standards as important contributors to the dramatic increase in foreclosures. Additionally, the difficulty of modifying mortgages and approving sales for a price lower than the outstanding principal balance (short sales) became evident with the diffuse ownership of mortgages through MBSs. In 2006, the Chicago Mercantile Exchange started issuing home price futures contracts with values based on the Case-Shiller Index for select cities. Like any futures contract, these are agreements between sellers and buyers to exchange housing contracts in the future at predetermined prices. If at the future date the actual contract price is higher than the agreed-upon price, then the buyer of the futures contract profits, while if at the future date the actual contract price is below the agreed-upon price, then the seller of the futures contract profits. An owner of residential real estate can use futures contracts to protect or hedge against price risk. For example, if the market price of a house falls, then the owner has a drop in net wealth. However, if an individual had sold a home price futures contract, then he or she may have earned a profit that cancelled the loss of owning housing. The ability of individuals, investors, and developers to protect themselves against price risk had been very limited before the introduction of these S&P/Case-Shiller Home Price futures, and it remains to be seen how well used these futures contracts will be. Mark Bertus, Harris Hollans, and Steve Swidler (2008) show that the ability to hedge against price risk in Las Vegas is mixed.

An active literature has grown regarding housing and the labor market. Andrew Oswald (1996) shows that across countries, home ownership rates and unemployment rates are positively correlated. Jakob Munch, Michael Rosholm, and Michael Svarer (2006) find evidence that geographic mobility is decreased with home ownership, but home ownership appears to reduce overall unemployment. With more and more people having mortgage principals greater than the current market value of their houses, it seems that job mobility and wages will be negatively impacted.

## Conclusion

This chapter reviewed the theory of real estate markets, including the important relationship of real estate rental price to selling price. Location is a key determinant of real estate rental price, but every parcel of real estate is unique, and hedonic pricing can determine the value that users place on individual characteristics of real estate. Most real estate is purchased with borrowed funds, and there have been important changes in mortgage financing over the last decade. The government treatment of real estate through taxation and zoning or land use restrictions will continue to be an important policy topic in the future. But perhaps the most visible issue in the short run is the impact of the real estate boom and bust. Future research will undoubtedly try to determine whether government policy can effectively prevent or mitigate the effects of similar episodes in the future.

## References and Further Readings

Anas, A., Arnott, R., & Small, K. A. (1998). Urban spatial structure. *Journal of Economic Literature, 36*(3), 1426–1464.

Bertus, M., Hollans, H., & Swidler, S. (2008). Hedging house price risk with CME futures contracts: The case of Las Vegas residential real estate. *Journal of Real Estate Finance and Economics, 37*(3), 265–279.

Case, K. E., Shiller, R. J., & Quigley, J. M. (2005). Comparing wealth effects: The stock market versus the housing market. *B. E. Journals in Macroeconomics: Advances in Macroeconomics, 5*(1), 1–32.

DiPasquale, D., & Wheaton, W. C. (1996). *Urban economics and real estate markets.* Englewood Cliffs, NJ: Prentice Hall.

Federal Reserve. (2009). *Flow of funds accounts of the United States.* Available at http://www.federalreserve.gov/releases/z1/current/z1.pdf

Frew, D., & Wilson, B. (2002). Estimating the connection between location and property value. *Journal of Real Estate Practice and Education, 5*(1), 17–25.

Gaspar, J., & Glaeser, E. L. (1998). Information technology and the future of cities. *Journal of Urban Economics, 43*(1), 136–156.

Glaeser, E. L., & Gyourko, J. (2003, June). The impact of building restrictions on housing affordability. *Economic Policy Review,* 21–29.

Glaeser, E. L., & Shapiro, J. M. (2003). The benefits of the home mortgage interest deduction. *Tax Policy and the Economy, 17,* 37–82.

Gould, E. D., Pashigian, B. P., & Prendergast, C. J. (2005). Contracts, externalities, and incentives in shopping malls. *Review of Economics and Statistics, 87*(3), 411–422.

Green, R. K. (2007). Airports and economic development. *Real Estate Economics, 35*(1), 91–112.

Haines, C. L., & Rosen, R. J. (2007). Bubble, bubble, toil, and trouble. *Economic Perspectives, 31*(1), 16–35.

Hamilton, J. (2005). What is a bubble and is this one now? *Econbrowser.* Retrieved December 28, 2008, from http://www.econbrowser.com/archives/2005/06/what_is_a_bubbl.html

Joint Committee on Taxation. (2008). *Estimates of federal tax expenditures for fiscal years 2008–2012* (JCS-2-08). Washington, DC: Government Printing Office.

Legislative Analyst's Office. (2007). *Tax expenditures reviews.* Available at http://www.lao.ca.gov/2007/tax_expenditures/tax_expenditures_1107.aspx

Mankiw, N. G., & Weil, D. N. (1989). The baby boom, the baby bust, and the housing market. *Regional Science and Urban Economics, 19,* 235–258.

Mayer, C. J., Pence, K. M., & Sherlund, S. M. (2008). *The rise in mortgage defaults* (Finance and Economics Discussion Series Paper No. 2008-59). Washington, DC: Board of Governors of the Federal Reserve System.

Munch, J. R., Rosholm, M., & Svarer, M. (2006). Are homeowners really more unemployed? *Economic Journal, 116,* 991–1013.

Muth, R. F. (1960). The demand for non-farm housing. In A. C. Harberger (Ed.), *The demand for durable goods* (pp. 29–96). Chicago: University of Chicago Press.

Oswald, A. (1996). *A conjecture of the explanation for high unemployment in the industrialised nations: Part 1* (Research Paper No. 475). Coventry, UK: Warwick University.

Shiller, R. J. (2007). *Understanding recent trends in house prices and home ownership* (NBER Working Paper No. 13553). Cambridge, MA: National Bureau of Economic Research.

U.S. Census Bureau. (2008, September). *American housing survey for the United States: 2007.* Available at http://www.census.gov/prod/2008pubs/h150-07.pdf

U.S. Census Bureau, Housing and Household Economic Statistics Division. (2009). *Housing vacancies and homeownership.* Available at http://www.census.gov/hhes/www/housing/hvs/historic/index.html

Voicu, I., & Been, V. (2008). The effect of community gardens on neighboring property values. *Real Estate Economics, 36*(2), 241–283.

Worzala, E., & McCarthy, A. M. (2001). Landlords, tenants and e-commerce: Will the retail industry change significantly? *Journal of Real Estate Portfolio Management, 7*(2), 89–97.

# 60

# Economics of Wildlife Protection

Rebecca P. Judge
*St. Olaf College*

Economics arrived late to wildlife protection. Although optimal timber harvest rates were first described by Martin Faustmann in 1849, and although Harold Hotelling had defined the parameters for profit-maximizing extraction of exhaustible resources like coal and petroleum by 1931, the problem of wildlife protection was ignored by the discipline until 1951. That year, Charles Stoddard (1951), a consulting forester working out of Minong, Wisconsin, urged fellow attendees at the North American Wildlife Conference to address the nation's "steady deterioration of wildlife habitat " with "an intensive development of a branch of knowledge devoted to the economics of wildlife management" (p. 248). The problem, according to Stoddard, was that wildlife production on private lands competed with more remunerative sources of income to landowners. For wildlife to recover, private landowners needed to be "provided with incentives, economic and otherwise, for producing wildlife crops" (p. 248).

But the discipline was slow to pick up on Stoddard's call. Although others joined in calling for public assistance to private landowners to promote increased investment in wildlife habitat on private lands, the economic problem of the optimal provision of wildlife remained largely ignored. The basic structure of the problem was obvious: Wildlife production on private lands competed with other uses, such as crop and timber production, for which the landowner received compensation. Therefore, although wildlife was certainly valued by society, the fact that private landowners were not reimbursed for their contributions to wildlife production meant that the supply of wildlife would be less than optimal. Following the economic logic first described by Arthur Pigou, wildlife production represented another case in which the marginal social value of a good exceeded its marginal private value; the value of wildlife to society was greater than its value to the private landowner. Pigou (1932) notes that in response to these situations, "it is theoretically possible to put matters right by the imposition of a tax or the grant of a subsidy" (p. 381) that would, in effect, equate private returns with social returns. Stoddard's (1951) call for public subsidies to private landowners to increase wildlife production was a request for a Pigouvian subsidy for wildlife.

Although Pigou's work was widely accepted as having laid the theoretical foundation for government intervention in private markets, it took another 30 years for economists to overcome their reticence to use these theoretical insights to argue for public policies that, while benefiting some, would impose costs on others. Seminal contributions by Harold Hotelling, J. R. Hicks, and Nicholas Kaldor allowed the discipline to arrive at a basic consensus that economists could recommend policies that would change the distribution of income in society so long as the gains experienced by some members of society as a result of the policy were greater than the losses experienced by others. This allowed economists to return to Pigou's work, using it to ground an argument for the use of subsidies and taxes to promote social welfare. Practitioners of the new welfare economics set about describing the conditions that would determine when such intervention was warranted.

Social and private benefits differ when the production or consumption of a good creates externalities. An externality exists when an economic decision generates tangible benefits or costs that are not received or incurred by the economic decision maker but rather by some third party. In the case of wildlife, private landowners bear many of the costs associated with wildlife production. These costs may be deliberately incurred, as when a private landowner forgoes

financially remunerative uses of the land to provide more suitable wildlife habitat, or they may be the unintended result of wildlife-related crop damage. Although private landowners bear the costs of wildlife production, they are seldom able to capture all of the benefits generated by wildlife. These benefits are instead enjoyed by hunters, hikers, bird-watchers, and other wildlife enthusiasts, as well as those who provide products like binoculars, camping gear, rifles, and vacation cabins to the wildlife enthusiasts. These benefits exist as externalities, and as such, they do not inform the landowner's land use decision.

When externalities exist, the resulting supply of the externality-producing good will differ from that associated with the maximization of net social benefits; the invisible hand of the free market will not lead all to the socially optimal result. In one of the earliest economic studies describing the value of wildlife, Hays Gamble and Ronald Bartoo (1963) estimated that the deer herd existing in Sullivan County, Pennsylvania, cost each farm in the county an average of $200 per year in crop damages, while simultaneously generating gross revenues to the county of about $546 per farm as a result of expenditures made in the county by hunters visiting the area. Although the county clearly had an interest in promoting the deer supply to attract more hunters to the area, the farmers, on whose land the herd depended, had an interest in reducing the herd's size. The presence of externalities caused the socially desirable outcome to deviate from that which occurred absent market intervention.

Despite the attractiveness of the Pigouvian argument, not all economists agree that externalities provide a sufficient rationale for government intervention in the market. These economists reject the Pigouvian solution of government taxes and subsidies in favor of a laissez-faire response best articulated by Ronald Coase. According to Coase and his followers, not only is the government's response likely to cause inefficiencies, but furthermore, government intervention is itself unnecessary. Absent government interference, the private parties affected by externalities and those creating the externalities will come to their own efficient solutions if the economic impact of the externality is sufficiently large. In the words of Harold Demsetz (1967), "Property rights develop to internalize externalities when the gains of internalization become larger than the cost of internalization" (p. 350). If the benefits of wildlife are significant enough, those who value the wildlife will do what it takes to see to it that wildlife is supplied. They can buy up land and convert it to wildlife reserves, or they can pay landowners to switch from crop production to wildlife production in exchange for the rights to enjoy the wildlife produced. According to the Coasians, if the benefits generated by the externality are not significant enough to warrant the internalization of these externalities, it follows that they are not significant enough to warrant government attention.

Although a discussion of the relative merits of the two positions is beyond the scope of this chapter, it is nonetheless a critical debate to those interested in the economics of wildlife protection because it informs how one both frames and addresses the question of wildlife supply. According to the Pigouvians, government response to the chronic undersupply of wildlife requires subsidization to private landowners for those actions that contribute to the supply of wildlife and taxation of private landowners for those actions that diminish wildlife supply. This policy recommendation stands in sharp contrast to that resulting from a Coasian approach to wildlife protection. In this case, government should see to it that the property right—to the unrestricted use and enjoyment of private land or to a population or species of wildlife—is clearly established among private parties, and then the parties should be encouraged to negotiate private solutions.

The debate between the Pigouvians and the Coasians remains unresolved and has resulted in a lack of coherency in the policy recommendation offered by economics for wildlife and other goods (or bads) created as externalities. Given the lack of consensus among economists, the government met no concentrated opposition to its own particular approach to wildlife protection, which was to acquire even more land for wildlife habitat and refuge and to use legislation to prohibit activities on public or private lands that threatened to depress wildlife populations to the level of possible extinction. Between 1964 and 1994, land holdings of the U.S. Fish and Wildlife Service increased nearly fourfold, growing from 22.7 million acres to 87.5 million (*Federal Lands,* 1996). In 1973, Congress passed amendments to the Endangered Species Act of 1973, making it illegal, on public or private land, "to harass, harm, pursue, hunt, shoot, wound, kill, trap, capture, or collect" (Section 3) any species found to be in danger of extinction. Rather than solving the externality problem by creating habitat conservation incentives, via taxes and subsidies to private landowners, or staying out of the wildlife problem altogether, the government approached the problem of declining wildlife populations by making wildlife protection and production one of its responsibilities. By the end of the 1960s, the role of economics in wildlife protection had changed from debating the proper management tool to measuring and describing the extent to which the government's own policies had the potential to improve overall social welfare. With few exceptions, the economics of wildlife protection moved from policy prescription to the careful description, characterization, and quantification of benefits of government-provided wildlife conservation efforts.

## Theoretical Foundations of Wildlife Protection: Wildlife and the Problem of Market Failure

Wildlife presents the economist with a near-perfect storm of exceptions to Adam Smith's argument that the invisible hand of individual self-interest will guide society to welfare-maximizing allocations of goods and services. Indeed, at least three characteristics of wildlife interfere with its

optimal allocation. Economic theory predicts that because of (1) the particular characteristics of the production of wildlife benefits, (2) the particular characteristics defining access to these benefits, and (3) the particular characteristics of the benefits themselves, the market-generated supply of wildlife will fall short of the optimum. This makes wildlife supply and wildlife protection an interesting and perplexing problem for economists.

## Market Failure in the Production of Wildlife Benefits

Wildlife is an ephemeral resource. Its nature is such that the very act of ownership transforms it to some other species of good. Thus, a wild deer becomes, on the act of extraction that is a precondition to ownership, someone's source of meat or potential trophy, pet, or unit of livestock. To be wildlife, or ferae naturae, is to be unowned and unownable. Once an animal is owned, it is no longer termed *wild.*

This concept, this metaphysical truth about wildlife, goes back at least as far as Roman law, which recognized wildlife as *res nullius,* literally, nobody's property. Wildlife differed in the Roman tradition from other natural resources, like water, which the Romans categorized as *res communis,* or common property, the property of everyone. The Roman government could claim for its people a property right to the waters within the empire's boundary, and it could protect that claim with force as necessary. It could not similarly claim for its people a property right to the wildlife that crossed international borders of its own will and according to its own tastes. Individuals could extract wildlife from the natural population to eat or domesticate, but that act of extraction changed the very nature of the extracted animal from *res nullius* to the property of the captor. Ownership removes the *wild* from *wildlife.*

The profound difference that exists between wildlife and owned animals constrains wildness and wildlife to exist as an externality. An animal cannot remain wildlife if it is produced as part of a deliberate breeding and management program. If a herd of buffalo exists as part of nature's endowment, it is wildlife, but if the herd exists as part of a rancher's assets, it is simply livestock. The valuable but ephemeral asset of wildness is diminished by the act of husbandry. Aldo Leopold (1986) noted, "The recreational value of a head of game is inverse to the artificiality of its origin, and hence in a broad way to the intensiveness of the system of game management which produced it" (p. 395). Perversely, attempts to address the externality of the wildlife benefit through the private management of the wildlife supply diminish the magnitude of the benefit itself.

Although the value attributable to the wildness of wildlife seems destined to exist as a nonexcludable externality generated by wildlife production, certain wildlife-associated benefits, such as those generated through the acts of hunting or viewing wildlife, are potentially excludable, because private landowners can limit entry to their private property. But even these excludable benefits are not likely to generate market compensation commensurate with the benefits generated. As a fugitive resource, wildlife does not necessarily reward those who practice good habitat management on their privately held land. Several characteristics of the wildlife production function, such as the magnitude of the land area typically involved in wildlife production, the fact that breeding habitat is often quite distinct and distant from grazing habitat, and the existence of chronobiologic phenomena like migration, mean that investments in wildlife-augmenting activities are unlikely to be able to capture the appropriate monetary recompense. If the wildlife population breeds and feeds on one parcel of land yet spends the hunting season on some other parcel, the owners of the breeding grounds cannot capture in hunting receipts the value of their contributions to wildlife production. Meanwhile, the owner of the prime hunting land potentially reaps an unearned profit. The inability to capture the benefits associated with the supply of wildlife means that wildlife will be undersupplied.

## Market Failure and the Property Regime Controlling Access to Wildlife

Many wildlife populations exist under conditions of open access, which is to say their range is so broad or unpoliced that anyone wishing to exploit them for personal gain can do so. Open access resources are susceptible to overexploitation because those who employ the resource have no incentive to exercise restraint in doing so; like children grabbing up candy after the piñata has disgorged its contents to the floor, those who depend on an open access resource know that moderation receives no reward. Although the rational individual owner under a system of private property will extract only to the point where profits are maximized (i.e., to the point at which the marginal revenue generated by the extracted resource is equal to the marginal cost of extraction from the resource) and although the tribe or community extracting from a commonly held resource may develop means of regulating its use to preserve its benefit stream, exploitation under conditions of open access will continue so long as any positive profits remain to be earned from the resource. This is because profits, or net benefits, generated by extraction from the wildlife population will continue to lure others into the same enterprise so long as those profits exist. Thus, there is a greater rate of exploitation for a given population of wildlife under conditions of open access than would exist if access to the wildlife resource could be controlled as private property or as a regulated commons.

H. Scott Gordon (1954) first described this phenomenon with respect to the ocean fishery. Noting that "fishermen are not wealthy, despite the fact that the fishery resources of the sea are the richest and most indestructible available to man" (p.132), Gordon went on to explain the low returns to fishermen by considering how a profitable fishing grounds will continue to attract new entrants until

such time as the potential entrants see no profit in exploiting the fishery. That is, entry continues into the fishery until all economic profits are exhausted. This is because those exploiting an open access resource have no reason to limit their extraction to promote the continued health of the fishery, because the benefits of any restraint will just go to some other fisherman. According to Gordon, "He who is foolhardy enough to wait for its proper time of use will only find that it has been taken by another" (p. 135). Open access explains the destruction of the North American buffalo herds, the near extinction of certain species of whales, and the precarious status of many species of African wildlife. The supply of a good existing under conditions of open access will be less than is economically efficient.

But even if access is limited to a tribe, licensed harvesters, or some other identifiable and closed group, collective ownership of a resource can lead to overexploitation, as Garrett Hardin (1968) describes in his paradigm-shifting article, "The Tragedy of the Commons." Overexploitation of the commons is the potential result of an asymmetry in the distribution of benefits and costs among those using a common property resource. Specifically, those who use the commons for their own purposes are able to capture for themselves the entire benefit associated with the use of the commons, while the costs that use imposes are shared by all who hold title to the commons. In the case of depredations on wildlife, although the hunter or trapper is able to gain the full benefit associated with capturing his or her prey (presumably by either selling the hide, consuming the carcass, or mounting its head on the library wall), insofar as extraction reduces the size of the wild population and requires, in turn, increased effort to extract the next animal from the population, this cost is divided among all who prey on the wild population. As such, the cost an individual bears when extracting a unit of resource from a commonly held population is less than he or she would bear as a result of extracting the same unit from a privately held population. The lowered perceived costs incurred by the individual extracting from a commonly held resource therefore results in more extraction for the same benefit function than would occur with a privately held property.

This is not to say that private ownership or management of wildlife would solve the problem of extinction. Indeed, as originally pointed out by Colin Clark (1973), extinction can be the efficient outcome for a renewable resource held by a profit-maximizing private owner. If revenues from exploiting a wild population (via hunting, fishing, or other extractive activity) exceed the costs associated with this effort and if the rate of growth of the species is less than the rate of return the owner could earn in some other investment, the owner has the economic incentive to convert the resource into cash to invest in the higher yielding asset. In other words, if the price a hunter can get for the last elephant tusk taken from the last elephant is greater than the costs the hunter incurs finding and killing this last elephant and if the rate of interest the hunter could earn on the net revenues from the sale of that tusk is greater than the rate of growth of the elephant population, the profit-maximizing hunter would find it in his or her interest to kill that last elephant. Private ownership might result in efficient management of the wildlife resource, but it offers no guarantee of the resource's preservation.

## Market Failure and the Benefits of Wildlife

The final factor contributing to the undersupply of wildlife concerns the public-good nature of many of the benefits wildlife generates. Public goods have two distinguishing characteristics. First, they are nonrival in consumption, meaning that more than one consumer may enjoy the same public good simultaneously. This differs from the rival nature of private good benefits, which can be enjoyed only by the single consumer possessing the good itself. The nonrival nature of the benefits generated by public goods means that a single unit of a public good bestows benefits simultaneously on all the consumers who are simultaneously consuming it. As a result, even if the benefits each individual derives from the provision of a unit of public good are very small, that unit can nonetheless contribute a very large amount to social welfare when its benefits are summed up over millions of individual consumers.

The other distinguishing characteristic of public goods is that their benefits are nonexcludable. Once a public good has been produced, it is impossible to restrict its consumption to a select group of individuals. This means that a public good cannot be rationed to those who pay for it. The inability to charge individuals for the use of public goods means that private producers will have no incentive to supply public goods to the market, because potential consumers will simply free ride and enjoy the goods' benefits without paying for them. Because those who do not pay for a public good enjoy the same level of benefit from the public good as those who do pay, public goods are unlikely to be produced and exchanged in a market economy. Government is necessary to ensure the production of public goods.

The public-good nature of the wildlife benefit entered the economic spotlight when John Krutilla (1967) published "Conservation Reconsidered." This seminal article changed forever the way economists, policy makers, and the general public characterized the services provided by environmental resources. Although economists since John Stuart Mill have used the existence of public goods to justify government provision of everything from lighthouses to national defense, Krutilla's essay expanded the list of public goods to include "some unique attribute of nature . . . a threatened species, or an entire ecosystem or biotic community essential to the survival of the threatened species" (p. 777). Krutilla based his argument on the assertion that the knowledge of the mere existence of these wild

resources constitutes "a significant part of the real income of many individuals" (p. 778). By including the existence of a threatened species within the genre of public goods, Krutilla made the strongest economic case yet for the preservation of species and the conservation of wildlife habitat. If indeed the actuality of a species provided millions of people with some form of income, then that species' existence had a tangible economic value that would be irretrievably lost should society choose to allow development to proceed to the point that it resulted in the species' extinction.

Krutilla (1967) describes the basic contours of the benefits emanating from preservation of unique environmental assets as equal to the "minimum which would be required to compensate such individuals were they to be deprived in perpetuity of the opportunity to continue enjoying the natural phenomenon in question" (pp. 779–780). In the case of extinction, this compensation would need to reflect the amount individuals would have been willing to pay to retain an option to use or interact with the species, the expected value of what might have been learned or extracted from the species had it not gone extinct, the value individuals had placed on the species' contribution to "the mere existence of biological . . . variety" (p. 782), and the value of the benefit individuals would have derived from leaving the species as a bequest to their heirs. Because all of these benefits are nonrival and nonexcludable, the preservation of any particular species of wildlife represents a public good of potentially enormous economic value that nonetheless is not and cannot be reflected in market exchanges.

Krutilla's (1967) expansive view of wildlife benefits changed wildlife from prey to be managed for hunters to a resource capable of providing enjoyment to an entire nation, even the entire world. The recognition of the public-good benefits resulting from the mere existence of a species of wildlife provided needed justification for efforts to preserve endangered species and populations of wildlife across the globe. Wildlife protection after "Conservation Reconsidered" enlarged to include endangered species preservation as one of its key concerns.

## Applications and Empirical Evidence

Economic solutions to the problems facing the supply of wildlife are largely limited to proposals of remedies for those attributes of the wildlife benefit and production functions that cause the market to fail to arrive at an efficient allocation of the wildlife good. That is, economists start by characterizing wildlife protection as an economic problem and then proceed to offer an economic solution to the problem. Although this seems to confirm Abraham Maslow's maxim that if one has only a hammer, one sees every problem as a nail, it also reflects the discipline's insistence on staying within its own area of expertise and imposing on itself its own standards for what constitutes advances in knowledge.

Some problems posed by wildlife are more amenable to economic solutions than others. For example, if wildlife is undersupplied because the market does not provide adequate compensation to those who incur the costs of wildlife production, a solution becomes both obvious and easy to implement, particularly for those benefits, like hunting and fishing, from which it is at least possible to exclude those who fail to pay.

But for other problems, the role of the economist in resolving issues pertaining to wildlife protection is less certain. It is one thing to point out the public-good nature of many of the benefits of wildlife protection; it is another thing entirely to arrive at estimates of preservation benefits and costs that allow the economist to describe for a given biome, ecosystem, or planet the optimal amount of preservation and extinction.

This section considers how economics has been used as a tool for informing optimal levels of wildlife protection. It begins with a review of economic solutions to the wildlife externality problem and continues to consider economic contributions to the understanding of how property regimes influence wildlife supply. This is followed by a description of how economics has contributed to both the understanding and the measurement of benefits flowing from wildlife. The section concludes with an examination of how economics has informed the question of endangered species preservation.

### Addressing Market Failure in the Production of Wildlife: A Look at the Pigouvian and the Coasian Solutions

Even as the economics profession was slow to respond to Stoddard's (1951) call to develop a branch of the discipline devoted to the economics of wildlife protection, so too was Congress sluggish in its response to the call for subsidies to private landowners to encourage wildlife production. Indeed, over 30 years elapsed between Stoddard's article and congressional legislation authorizing subsidies to landowners for habitat conservation efforts. The 1985 farm bill, known as the Food Security Act of 1985, established a Conservation Reserve Program (CRP) that offered farmers rental payments of about $50 per acre per year for agricultural land taken out of crop production (Section 1231). Although initially designed as both a means of preventing soil loss on highly erodible land and a way to increase commodity prices through commodity supply reductions, the program is now viewed by the U.S. Department of Agriculture (USDA, 2007) as a primarily environmental program. Nearly 40 million acres of agricultural land are presently enrolled in the program, although the 2008 farm bill has reduced the target for future enrollments to 32 million (USDA, 2009). Additional

landowner subsidy programs have been added to the mix—the Wetlands Reserve Program, the Environmental Quality Incentives Program, the Conservation Security Program, the Wildlife Habitat Incentives Program, and the Grasslands Reserve Program—which together add another 21 million acres of privately held land to those CRP lands managed with the goal of promoting wildlife habitat, according to the USDA. The total cost of these programs was estimated at $4.961 billion in 2008 (Cowan & Johnson, 2008).

The rationale for these subsidies, as introduced in the 1985 farm bill, sounds distinctly Pigouvian. Section 1231 of the 1985 legislation calls on the secretary of agriculture to enter into contracts with private landowners to encourage them "to conserve and improve the soil, water, and wildlife resources" (Food Security Act of 1985). However, insofar as the government makes no attempt to set the level of these subsidies to reflect the difference between the social and private welfare generated by conservation activities on private land, the subsidies fall far short of Pigou's suggested instrument. The informational demands of marginal benefits–marginal costs analysis make it practically impossible to estimate the requisite level of the Pigouvian subsidy, let alone arrive at an efficient means of taxing individuals for the benefits such a subsidy would generate. For this reason, Nobel Laureate James Buchanan (1962) criticized the Pigouvian solution as "not only politically unimaginable in modern democracy: it is also conceptually impossible" (p. 27).

Rather than tackling the unimaginable and impossible, applied economic research on subsidies to promote wildlife production has concentrated on evaluating the efficacy of the existing set of government programs. Although the CRP initially made no attempt to target cost-effective means of achieving conservation goals, later enrollments in the CRP have considered the ratio of environmental quality benefits to their costs, with salutary effects on program cost effectiveness (Osborn, 1993). However, because wildlife benefits appear to be uniformly distributed across CRP-eligible land, improving the overall cost effectiveness of the CRP enrollment appears to have little impact on the wildlife benefit generated by CRP participation (Babcock, Lakshminarayan, Wu, & Zilberman, 1996). Recent research on the cost effectiveness of government conservation subsidy programs compares the relative efficacy of land retirement programs like the CRP to programs like the Environmental Quality Incentives Program that keep the land employed while mandating landowner investment in ecosystem improvement. Here, the published evidence is somewhat mixed. Some (Polasky, Nelson, Lonsdorf, Fackler, & Starfield, 2005) have found the potential for significant savings is realized by satisfying a wildlife conservation goal through a mixed land use strategy that provides both wildlife habitat and crop production. Still others have obtained results that indicate that complete and total land retirement is a more cost-effective means of conserving wildlife (Feng, Kurkalova, Kling, & Gassman, 2006). Clearly, this is an area where more research should prove both enlightening and productive.

Given the necessarily wild nature of wildlife, attempts to implement a Coasian solution to the wildlife undersupply problem are handicapped by an inability to assume away the problem by assigning property rights to the wildlife resource. The metaphysical essence of wildlife prevents it from becoming someone's property or herd. Coasian-inspired research has therefore focused on identifying the incentives that contribute to changes in the overall size of the wildlife population. These incentives can be directed at those whose actions affect the supply of wildlife as well as those responsible for changes in the demand for wildlife.

Regarding wildlife supply, Coasian-inspired analysis has examined how landowners respond to direct payments to create more wildlife habitat. Not surprisingly, research reveals consistently that landowners whose only interest in land is as a productive input to some market-related activity require higher payments to engage in habitat production than landowners whose interests in the land are not purely financial. Therefore, the somewhat expected result emerging from the analysis of landowner willingness to participate in wildlife conservation programs is that motives matter. Those landowners who exhibit an interest in wildlife or are not dependent on the land for their livelihoods require lower incentive payments than those who report no interest in wildlife or have only a financial interest in their own land holding. Whether this result would hold if habitat payments become standard across landowners is, of course, a matter of speculation. If government were to institute a large-scale program to pay landowners for their contributions to wildlife habitat, it could use its monopsonistic powers to discriminate among landowners based on these differences in wildlife habitat supply elasticity.

Coasian-inspired analysis of those factors that contribute to demand for the extractive use of wildlife focuses on the efficacy of trade bans as a means of reducing extractive pressure on a wildlife population. The desirability of the trade ban has been shown to depend on the ability of the state to protect the wildlife resource from illegal predations. If the police powers of the state are so extensive that the expected returns to poaching, given the probability of arrest, fines, or detention, are less than many would-be poachers could earn in their next-best income-generating opportunity, then trade bans should be discouraged because they reduce the returns to state-sponsored wildlife stewardship. That is, the state's incentives to protect its threatened wildlife are decreased under a trade ban, because the ban limits the potential for remuneration from the harvest of the protected species. On the other hand, if the police powers of the state are so nugatory that the species exists under what are essentially conditions of open access, import bans might be desirable

because they lower the potential returns from exploiting the wildlife resource by restricting the size of its potential market. Mixed and muddled empirical results seem to confirm the theoretical ambiguity regarding the desirability of these bans. While Pigouvian subsidies will continue to be subjected to efficacy tests, Coasian attempts at limiting government intervention in wildlife exchanges and interactions will confront continued tests of their overall desirability.

## Addressing Market Failure in the Access to Wildlife: Understanding the Role of Property Regimes

Perhaps the most well-known application of economic theory to address the wildlife access problem concerns the development and implementation of individual transferable quotas (ITQs). Like a license, which represents one way of regulating access under a common property regime, an ITQ is a government-granted right to extract a certain amount of a resource. But unlike a license, the individual possessing the ITQ can sell this right to another. By allowing sales, ITQs provide their owners with a means of capturing some of the benefits of conservation efforts. As the wild population increases in size, the ITQ increases in value because presumably less effort need be expended to harvest the same amount of resource. The ITQ addresses the open access problem by restricting access to the resource to only those individuals who possess licenses, and it attempts to solve the potential common property problem by rewarding restraint in resource extraction through increases in the value of the ITQ.

Unlike some economic policy suggestions, ITQs have an observable track record. New Zealand adopted the ITQ system as a means of regulating its fishery in 1986. A similar system was implemented 5 years later in Iceland. Studies evaluating the ITQ system give it relatively high marks for promoting economic efficiency. As predicted, the system allows the same amount of the resource extraction with less effort than is exerted under other systems regulating the commons (Arnason, 1991; Yandle, 2008). However, ITQs have admittedly undesirable distributional consequences, because those able to buy up ITQs can effectively close a regional fishery and thereby cause localized economic collapse and dislocation (Eythórsson, 2003).

Other empirically based research examines the effect of alternative property regimes on wildlife supply. Relying primarily on case studies in lieu of theoretical abstraction, the litmus test applied to the various property regimes is typically not whether a given ownership configuration is likely to promote economic efficiency so much as whether the ownership pattern will conserve the wildlife resource (Brown, 2000, p. 902). As such, these studies reflect what for some is a paradigm change in the economics of wildlife protection, because sustainability of the wildlife resource replaces efficiency as the norm about which property regimes are to be judged.

## Addressing Market Failure and the Benefits of Wildlife

Convinced by Krutilla's (1967) argument that unique natural assets generate more benefits than are revealed through market transactions, economics developed a distinct typology of those benefits generated through interaction with the natural environment. Within this classification system, benefits sort into two main categories: use values and nonuse values. Those benefits generated as the result of direct in situ interaction with an asset are use values, while nonuse values are those that require no direct interaction with the environmental amenity. Hunting, bird-watching, nature photography, and fishing are all use values, while knowledge—of the existence of a wildlife population or a scenic wonder like the Grand Canyon—is a nonuse value.

The taxonomy of values includes a second tier of subordinate classifications. Use values can be consumptive or nonconsumptive, depending on whether they reduce the stock of the natural asset. Hunting and fishing represent consumptive uses of wildlife, while bird-watching and nature photography are nonconsumptive uses of the wildlife resource. Nonuse values are categorized on the basis of the type of knowledge generating the value. Knowledge that the resource exists to fulfill a potential future desire to interact with it contributes to the option value of the wildlife resource, while knowledge that the resource exists for the use and enjoyment of future generations contributes to its bequest value. Existence value is the monetized measure of satisfaction that one experiences from simply knowing that the wildlife population exists, independent of any desire to interact with it directly or save it as a bequest for one's heirs.

Benefit estimation techniques for these nonmarket goods differ according to whether one is estimating use or nonuse values. For a use value like bird-watching, even if the use itself is not allocated via markets, economists can estimate the value of the benefits it generates by observing how demand for complementary goods, like binoculars, changes with changes in the quantity of the nonmarket good consumed. From these changes, one can derive at least a minimum value for the nonmarket good. However, this technique is not available in the case of nonuse values, which we consume without leaving behind any "behavioral trail" (Smith, 1990, p. 875). Without market generated prices and quantities on which to base benefit calculations, economists have been forced to abandon their insistence on revealed preference and seek some other way of determining the value of these scarce goods that fail to exchange in markets. The contingent valuation method (CVM) is the invention to which this necessity gave birth.

Although initially a cause of great controversy within the discipline, CVM is now the orthodox method of quantifying nonuse values. In a CVM survey, individuals are asked to reveal their willingness to pay for the right to enjoy a particular environmental amenity or their willingness to accept payment to forgo the enjoyment of the amenity. Typically, a CVM survey describes an environmental asset and then proposes a scenario in which the quantity of the asset is changed. Survey respondents are asked to describe how much they value that change. Using individual responses to hypothetical changes in the quantity of an environmental asset, economists can estimate its value.

But even though CVM is now well accepted as a means of quantifying nonuse benefits, concerns remain. One persistent issue is whether to measure the respondents' willingness to pay (WTP) to acquire a certain amount of an environmental asset or to instead measure their willingness to accept (WTA) payment to forgo the use of the asset. Lest this seem like a merely semantic difference, consider the likely difference between the amount an individual might require from a cousin in exchange for the engagement ring a grandmother bequeathed to the individual and the amount the same individual would offer to purchase the identical ring from his or her cousin if the cousin had inherited the ring instead. Property rights matter. One's WTP is constrained by one's budget constraint, but no such constraint limits one's willingness to accept payment. As a result, WTA has been found to be higher, for certain goods, than WTP. But this is not solely a result of the effect of the budget constraint. Rather, even controlling for the income, the fewer the substitutes that exist for the good, the greater the difference between WTP and WTA. Although the price that would induce an individual to part with a grandmother's engagement ring might be significantly greater than what he or she would be willing and able to pay to purchase it from a cousin, one might expect that the price that would induce the same individual to part with his or her present refrigerator is similar to the price the individual would be willing to pay to purchase a similar one. There are many likely substitutes for a given refrigerator; there are no substitutes for a grandmother's engagement ring.

As concerns payments to conserve wildlife, contingent valuation surveys have evinced significant differences between WTP and WTA. Richard Bishop and Thomas Heberlein (1979) found that Wisconsin duck hunters required nearly five times more in payment to surrender their licenses than they were willing to pay to purchase the same license. David Brookshire, Alan Randall, and John Stoll (1980) found that elk hunters were willing to pay only one seventh as much to purchase a license from the state to hunt elk as they were willing to accept in payment for the retirement of a license that they had been given by the state. G. C. van Kooten and Andrew Schmitz (1992) discovered that Canadian landowners were willing to pay only $3.90 per acre annually for a permit entitling them to drain wetlands but needed $26.80 per acre in compensation from the government to refrain from draining wetlands. These observed discrepancies between WTP and WTA lead to a conclusion that private activities that reduce the amount of wildlife impose a greater cost on society than is reflected in the measured WTP that appears in many CVM studies. In the words of Jack Knetsch (1990), the observed differences between WTA and WTP mean that "it is likely that, among other implications, losses are understated, standards are set at inappropriate levels, policy selections are biased, too many environmentally degrading activities are encouraged, and too few mitigation efforts are undertaken" (p. 227). If landowners had to pay the public for activities they engage in that diminish the wildlife resource, fewer of these activities would take place.

Contingent valuation studies have provided a set of estimates of interest to wildlife protection. Measured WTP for individual species ranges from a low of $8.32 annually (2006 dollars) to avoid the loss of the striped shiner to $311.31 per year (2006 dollars) to support a 50% increase in the saltwater fish population of western Washington and the Puget Sound (Richardson & Loomis, 2009). International visitors to China have a measured average WTP of $14.86 (1998 dollars) to conserve habitat for the giant panda (Kontoleon & Swanson, 2003). Citrus County, Florida, residents appear to be willing to pay a combined total of $194,220 (2001 dollars) per year to protect the Florida manatee (Solomon, Corey-Luse, & Halvorsen, 2004). As concerns the preservation of rare and endangered species, estimated WTP has increased over time (Richardson & Loomis, 2009). Evidence exists that WTP to protect a species grows at an increasing rate with the degree of perceived threat and at a decreasing rate as the population of a threatened species increases (Bandara & Tisdell, 2005).

WTP estimates reveal interesting and novel insights into wildlife conservation motives. Pennsylvania duck hunters were willing to pay more to avoid a reduction in duck populations resulting from global climate change than they were to avoid the same population reduction when caused by agricultural practices (Kinnell, Lazo, Epp, Fisher, & Shortle, 2002). A survey of Dutch households found that the respondents were willing to pay more to prevent reductions in the native seal population caused by oil spills than by a lethal, naturally occurring virus (Bulte, Gerking, List, & de Zeeuw, 2005). Economic research on payments for changes in wildlife populations continues to chip away at the margins, providing insight to many individual cases and refining and codifying survey techniques and statistical methods of analysis.

Economics has addressed the problems posed by the public-good nature of many wildlife benefits by developing new tools to measure these benefits. Once quantified, these values have been able to influence public policy through their inclusion in benefit-cost analysis. As such,

although economics has not been able to solve the problem created by the existence of public goods, it has been able to propose partial remedies that are at least potentially welfare enhancing. Additionally, economic explorations into the differences between WTP and WTA have provided important insights into human behavior, even as they challenge the discipline to develop consistent theoretical explanations. Equally interesting and challenging for the discipline is the recent empirical evidence that WTP for wildlife conservation depends on the factors that make such payments necessary. Ducks threatened by agricultural practices are not the same as ducks threatened by climate change, and seals threatened by oil spills appear to be more valuable than seals threatened by a virus. Future research to explain these observed differences promises to both complicate and improve the discipline of economics.

## Other Economic Contributions to Wildlife Protection: Economics of Species Preservation

Although economics has made great strides in helping provide information of value to policy makers as they make discreet decisions regarding wildlife conservation, the discipline has had difficulty developing a coherent and useful response to questions concerning wildlife preservation policy and strategy. This is largely because the benefits of biodiversity remain unknown and largely speculative, even though all agree that biodiversity is a good—and a pure public good, at that. Furthermore, while CVM might be reasonably effective at elucidating individual WTP to preserve a particular species under conditions of partial equilibrium, it has been unable to determine a WTP to preserve biodiversity in general. CVM estimates can be used to inform preservation decisions on a solely case-by-case basis and not as part of some overall preservation policy.

Economic contributions to the question of how much biodiversity to preserve include the controversial approach of recommending that all species be managed to preserve a population of sufficient size to avoid the threat of extinction. This minimally viable population size, known as the *safe minimum standard* (SMS), was originally described by Siegfried von Ciriacy-Wantrup in 1952. Ciriacy-Wantrup (1968) argued that conservation requires a collective choice rule that subjects "the economic optimum to the restriction of avoiding immoderate possible losses" (p. 88). This SMS functions as "an insurance policy against serious losses that resist quantitative measurement" (Ciriacy-Wantrup & Phillips, 1970, p. 28). As insurance, investments in the SMS should themselves be constrained; the SMS should be adopted unless the social costs of doing so are unacceptably large (Bishop, 1978). The abandonment of marginal analysis and the lack of rigor inherent in a policy based on judgments of what constitutes immoderate losses and unacceptably large social costs inhibit widespread adoption of the SMS within the discipline of economics.

As a choice made under uncertainty, the SMS has been portrayed as a dominant strategy in a two-person game between society and nature. In this game, society chooses between development and preservation, while nature chooses whether to inflict a disease on society and whether to encode a cure for that disease within the species threatened with extinction. If the decision rule calls for minimizing the maximum possible loss, the game yields inconsistent results, depending on whether the source of the uncertainty is whether the disease outbreak will occur or whether the cure will be found in the species (Ready & Bishop, 1991). However, if the decision rule calls for society to minimize its regrets instead of its maximum possible losses, the game's results consistently favor the SMS. Under a regret-minimizing strategy, society makes choices that minimize the potential losses associated with being wrong. If society chooses development when it should have chosen preservation, the losses are greater than if society chooses preservation when it should have chosen development (Palmini, 1999). However, although regret minimization may well inform individual behavior, economists have not been able to make a case for regret minimization as an economic norm. Thus, the SMS may describe a human behavior but not an economic objective.

The SMS approach demands that the economic objective of welfare maximization be replaced by that of preservation whenever extinction is a real threat, and this demand inhibits its widespread adoption by the economics profession. By demanding a discrete policy regime shift from the pursuit of welfare enhancement to the pursuit of preservation, the SMS exposes itself to criticism of its inconsistency. Because the rationale for this shift has eluded any formal statement, the SMS enjoys only limited support as an economic norm. Applied economic research regarding the SMS has focused on measuring both the WTP for preservation and the threshold level for what constitutes excessive preservation costs, as revealed by empirical analysis of costs incurred in the preservation of endangered species.

At the heart of the dissatisfaction with the SMS is its inability to help inform the choice between preservation and development that is the core issue for endangered species policy. The SMS approach argues for the preservation of all species, and although this argument is certainly appealing and popular, it is not economics. As Gardner Brown (2000) wrote poignantly, "Economists know what many ecologists cannot bear to admit, that not all species can be saved in a world of scarce resources" (p. 908). Brown's remarks echo those of Andrew Metrick and Martin Weitzman (1998), who critically observe, "At the end of the day, all the brave talk about 'win-win' situations, which simultaneously produce sustainable development and conserve biodiversity, will not help us to sort out how many children's hospitals should be sacrificed in the name

of preserving natural habitats. The core of the problem is conceptual. We have to make up our minds here what it is we are optimizing. This is the essential problem confounding the preservation of biodiversity today" (p. 21).

In place of the SMS, some economists have argued that preservation decisions should maximize biodiversity, subject to a budget constraint, by giving marginal preference to those species with no close substitutes as measured by the number of species within a genus. According to this model, the next dollar of preservation investment is more productively spent on a species with no close substitutes than on one that shares its genus with many other species. Of course, this assumes both that society has decided that biodiversity is the good whose production it values most and that the system of Linnaean taxonomy, an arguably human construct, accurately reflects the sort of biodiversity that society wishes to maximize. Economists remain frustrated by the inability of their discipline to help inform one of the central debates of our times.

## Policy Implications

The immediate and obvious result of decades of economic inquiry into the challenges facing wildlife protection is that there is ample reason to believe that the present supply of wildlife is smaller than optimal and that government intervention is necessary to protect and augment it. Government intervention can be in the form of direct subsidies to private landowners, rearrangement of property regimes and the rules of access, trade bans, absolute prohibitions on activities that degrade the wildlife resource, and public acquisition of land and resources as necessary to augment the supply of wildlife.

Regarding government subsidies to private landowners providing habitat for wildlife, economic research indicates that the current practice of removing land from production might be improved by offering a mix of programs, some of which offer subsidies to encourage multiple uses of the land base. This promises to lower the costs of achieving some wildlife protection goals. As the relative cost effectiveness of these alternative land management regimes appears to differ with the particular characteristics of the landscape and its native wildlife, government should be prepared to tailor its incentive programs to arrive at the most cost-effective subsidy mix.

Government attempts to rearrange property regimes by providing a transferable right to wildlife extraction have succeeded in reducing the overall cost of the extractive effort and have thus promoted efficiency. However, although privatization of a publicly held good, like wildlife, might lead to more efficient exploitation of the resource, it might also lead to its extinction or to a smaller population size than would occur under a common property regime. If the goal of the property rights rearrangement is species conservation, there is little evidence supporting the efficacy of privatization as a means to realize this goal.

The overall impact of trade bans on endangered species populations has not been settled empirically. Because the trade ban reduces the market value of the traded species, it is not clear from the empirical evidence whether this in turn leads to reduced poaching or to reduced investment in species conservation. To be effective instruments of species preservation, trade bans might need to be coupled with direct payments to those bearing the costs of conservation efforts.

Both the policies resulting in absolute prohibitions on otherwise productive activities that threaten the wildlife stock and the use of government resources to promote conservation activities should be guided by a consideration of the benefits of these activities relative to their costs. In estimating the value of the benefits of the wildlife resource, policy makers need to be cognizant of the difference between WTP and WTA estimates of wildlife values. Insofar as wildlife belongs to the nation, reductions in wildlife are losses in the real income of a nation's citizens. Therefore, the correct measure of the loss is obtained by asking citizens how much they would be willing to accept as compensation for a reduction in the population level, or the extinction, of a species. However, if increasing the population of wildlife entails constraining actions on private property, respondents should be asked to consider how much they are willing to pay private property owners for the loss of a beneficial land use. Values obtained in this manner should be used to guide public appropriations to preservation efforts.

Although economics can be used to inform the level of welfare-enhancing preservation investment in individual species, public policy has eschewed using this information in its preservation decisions. Domestic endangered species policy is absolutist; the law requires that any species found to be in danger of extinction is to be protected, regardless of either the costs or the benefits associated with its preservation. This implies that society has embraced some version of the SMS, at least as a matter of public policy. However, empirical analysis of the listing process under the Endangered Species Act indicates that species protection is influenced by interest group participation in the rule-making process (Ando, 1999). Listing decisions have also been shown to be sensitive to the likely impact of species protection on the districts of key congressional stakeholders (Rawls & Laband, 2004). Policy makers wishing to depoliticize the preservation process might well consider including a consideration of the biodiversity benefits of a given preservation option as part of their deliberations prior to awarding protection to a species.

## Future Directions

Trends in population growth, habitat modification, energy use, and waste production signal that wildlife populations will face increasing pressure in the future. Expanded use of the land base to meet the consumption

needs of an ever-growing world population implies further that the opportunity cost of wildlife habitat is also likely to increase in the future. Finally, increasing political unrest in areas of the globe stressed by water shortages and climate change is likely to take its toll on native wildlife populations. In sum, the suboptimal supply of wildlife is likely to continue unabated and become more profound in the years to come.

Economics can contribute to ameliorating these problems with research concerning the most cost-effective means of providing wildlife habitat. This includes the development of spatially explicit landscape models linking conservation costs to the wildlife supply response across a variety of landscape types, thus allowing the manager to choose the least-cost method of achieving a particular conservation goal. Another line of research that holds promise for promoting conservation cost effectiveness examines incentives motivating landowner participation in conservation programs. Wildlife managers can use information gleaned from these studies to target those landowners most willing to participate in conservation efforts and therefore lower conservation costs.

An important area for new research involves developing mechanisms to reward those who engage in wildlife conservation efforts and compensate those who bear the costs of species preservation. This means identifying and addressing the perverse incentives to eradicate species that are created by preservation mandates that constrain traditional land uses and the use of private property. It also means continued research on the effects of divergent property regimes on conservation and wildlife protection.

Perhaps the most important contribution economics can make to the problems posed by wildlife protection is that of values discernment. Although existence values provide much of the economic basis for investment in wildlife preservation, they constitute a rather gross aggregation of the nonuse benefits accruing to individuals as a result of the preservation decision. These values need to be further refined and teased into their constituent parts if economists are to improve how they inform preservation decisions.

Although the public has consistently expressed both substantial interest in avoiding extinctions and measurable WTP to avoid them, economists need to develop better ways to help the public describe just what it is that they wish to avoid via these expenditures. The present disciplinary emphasis on maximizing biodiversity assumes that the public values genetic distance over, say, honoring an ethical commitment to stewardship or maintaining something wild in an increasingly tamed world. Would the public be less concerned with extinction if the science of cloning progresses to the point where animals can be recreated from saved genetic material? What distinguishes the value of a wild specimen from a particular species with an otherwise identical specimen living in a zoo? What are the valuable attributes of wildness? These are the questions that are likely to be addressed by future economics research, because these are the questions that need to be answered if preservation is going to be allocated by informed, transparent policy decisions.

Taken together, these future directions for wildlife economics imply a more interdisciplinary, more nuanced approach to research, which will increasingly rely on input from the natural sciences as well as other social sciences. Research teams including economists, biologists, psychologists, sociologists, political scientists, and soil scientists are likely to become both more common and more necessary as conservation and protection move beyond the wildlife reserve to lands devoted to other productive uses. Not only will economists increasingly rely on other disciplines in the conduct of their research, but they will also need to continue to improve upon methods of engaging in conversation with the public at large to elucidate the values informing the desire to engage in conservation efforts. Simple-minded, predictable Homo economicus will be replaced by sophisticated and complicated real human beings as the object of economic research on values and objectives for wildlife protection. How interesting and how ironic that in studying wildlife, economists end up learning more about being human.

## Conclusion

The economics of wildlife protection has struggled over the years to incorporate a fugitive, unowned, and unownable natural asset into a discipline that is most comfortable describing exchanges of discretely bound goods in a market economy. Indeed, the nature of wildlife as an economic good is so unique that the discipline's early work in the field was devoted almost entirely to taxonomic descriptions of the benefits generated by wildlife and the property regimes that govern human interactions with the wildlife resource. These efforts were rewarded by a greater understanding of both what economists mean when they speak of value and the role property regimes play in determining resource conservation and exploitation.

Both wildlife protection and economics have benefited as a result of the development of this field of economic research. To the field of wildlife protection, economics has brought an explicit recognition of opportunity costs and their impact on the decisions of private landowners faced with deciding between promoting habitat conservation or development. Economics has emphasized the importance of understanding how distribution of costs and benefits flowing from wildlife conservation and extractive activities affect ultimately the rate of extraction, even the probability of extinction, of a wild population. Economics has expanded the number of stakeholders explicitly recognized by wildlife managers to include those who receive real but largely unobservable nonuse values from the wildlife resource.

Engagement with the questions and issues posed by wildlife protection has also benefited economics. The challenges confronted when working with such an unconventional subject have forced economics to grow beyond

its regular confines. For a discipline accustomed to dividing the world into consumers and producers, the challenges posed when examining a system in which benefits can be enjoyed by millions, even billions, of persons, absent any act of consumption, and a system in which goods are supplied as part of a natural, replenishable endowment have proved both liberating and confounding. Disciplinary engagement with the problem posed by wildlife protection has revealed that welfare can be augmented or diminished independent of the act of consumption. Indeed, economics has shown that sometimes something as gossamer as the mere knowledge of a good's existence constitutes an important part of an individual's asset portfolio. This is a revolutionary concept for economics. Similarly, the idea of a supply function characterized by a range in which it irreversibly heads to zero has challenged the discipline to develop ways to consider whether and how to avoid treading into that range. This has in turn made the discipline more receptive to exploring alternatives to the private-property and market-oriented regimes as a means of attaining socially desired outcomes.

## References and Further Readings


Ando, A. W. (1999). Waiting to be protected under the Endangered Species Act: The political economy of regulatory delay. *Journal of Law and Economics, 42,* 29–60.

Arnason, R. (1991). Efficient management of ocean fisheries. *European Economic Review, 35*(2–3), 408–417.

Babcock, B. A., Lakshminarayan, P. G., Wu, J., & Zilberman, D. (1996). The economics of a public fund for environmental amenities: A study of CRP contracts. *American Journal of Agricultural Economics, 78*(4), 961–971.

Bandara, R., & Tisdell, C. (2005). Changing abundance of elephants and willingness to pay for their conservation. *Journal of Environmental Management, 76*(1), 47–59.

Bishop, R. C. (1978). Endangered species and uncertainty: The economics of a safe minimum standard. *American Journal of Agricultural Economics, 60*(1), 10–18.

Bishop, R. C., & Heberlein, T. A. (1979). Measuring values of extramarket goods: Are indirect measures biased? *American Journal of Agricultural Economics, 61*(5), 926–930.

Brookshire, D. S., Randall, A., & Stoll, J. (1980). Valuing increments and decrements in natural resource service flows. *American Journal of Agricultural Economics, 62*(3), 478–488.

Brown, G. M. (2000). Renewable natural resource management and use without markets. *Journal of Economic Literature, 38*(4), 875–914.

Buchanan, J. M. (1962). Politics, policy, and the Pigovian margins. *Economica, New Series, 29*(113), 17–28.

Bulte, E., Gerking, S., List, J., & de Zeeuw, A. (2005). The effect of varying the causes of environmental problems on stated WTP values: Evidence from a field study. *Journal of Environmental Economics and Management, 49*(2), 330–342.

Ciriacy-Wantrup, S. (1968). *Resource conservation: Economics and policies* (3rd ed.). Berkeley: University of California Press.

Ciriacy-Wantrup, S., & Phillips, W. (1970). Conservation of the California tule elk: A socioeconomic study of a survival problem. *Biological Conservation, 3,* 23–32.

Clark, C. (1973). The economics of overexploitation. *Science, 181*(4100), 630–634.

Cowan, T., & Johnson, R. (2008). *Agriculture conservation programs: A scorecard* (CRS Report No. RL32940). Washington, DC: Congressional Research Service.

Demsetz, H. (1967). Toward a theory of property rights. *American Economic Review, 57*(2), 347–359.

Endangered Species Act of 1973, Pub. L. No. 93-205, § 3, 87 Stat. 884 (1973).

Eythórsson, E. (2003). Stakeholders, courts and communities: Individual transferable quotas in Icelandic fisheries 1991–2000. In N. Dolsak & E. Ostrom (Eds.), *The commons in the new millennium* (pp. 129–168). Cambridge: MIT Press.

*Federal lands: Information on land owned and acquired: Testimony before the Subcommittee on Oversight and Investigations of the Senate Committee on Energy and Natural Resources,* 104th Cong. (1996). (Testimony of Barry Hill). Retrieved from http://archive.gao.gov/papr2pdf/156123.pdf

Feng, H., Kurkalova, L., Kling, C., & Gassman, P. (2006). Environmental conservation in agriculture: Land retirement vs. changing practices on working land. *Journal of Environmental Economics and Management, 52*(2), 600–614.

Food Security Act of 1985, Pub. L. No. 99-198, § 1231, 99 Stat. 1354 (1985).

Gamble, H., & Bartoo, R. (1963). An economic comparison of timber and wildlife values on farm land. *Journal of Farm Economics, 45*(2), 296–303.

Gordon, H. (1954). The economic theory of a common-property resource: The fishery. *Journal of Political Economy, 62*(2), 124–142.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*(3859), 1243–1248.

Kinnell, J., Lazo, J., Epp, D., Fisher, A., & Shortle, J. (2002). Perceptions and values for preventing ecosystem change: Pennsylvania duck hunters and the Prairie Pothole region. *Land Economics, 78*(2), 228–244.

Knetsch, J. (1990). Environmental policy implications of disparities between willingness to pay and compensation demanded measures of values. *Journal of Environmental Economics and Management, 18*(3), 227–237.

Kontoleon, A., & Swanson, T. (2003). The willingness to pay for property rights for the giant panda: Can a charismatic species be an instrument for nature conservation? *Land Economics, 79*(4), 483–499.

Krutilla, J. (1967). Conservation reconsidered. *American Economic Review, 57*(4), 777–786.

Leopold, A. (1986). *Game management.* Madison: University of Wisconsin Press.

Metrick, A., & Weitzman, M. (1998). Conflicts and choices in biodiversity preservation. *Journal of Economic Perspectives, 12*(3), 21–34.

Osborn, T. (1993). The Conservation Reserve Program: Status, future, and policy options. *Journal of Soil and Water Conservation, 48*(4), 271–278.

Palmini, D. (1999). Uncertainty, risk aversion, and the game theoretic foundations of the safe minimum standard: A reassessment. *Ecological Economics, 29*(3), 463–472.

Pigou, A. (1932). *The economics of welfare* (4th ed.). London: Macmillan.

Polasky, S., Nelson, E., Lonsdorf, E., Fackler, P., & Starfield, A. (2005). Conserving species in a working landscape: Land use with biological and economic objectives. *Ecological Applications, 15*(4), 1387–1401.

Rawls, R., & Laband, D. (2004). A public choice analysis of endangered species listings. *Public Choice, 121*(3–4), 263–277.

Ready, R., & Bishop, R. (1991). Endangered species and the safe minimum standard. *American Journal of Agricultural Economics, 73*(2), 309–312.

Richardson, L., & Loomis, J. (2009). The total economic value of threatened, endangered and rare species: An updated meta-analysis. *Ecological Economics, 68*(5), 1535–1548.

Smith, V. (1990). Can we measure the economic value of environmental amenities? *Southern Economic Journal, 56*(4), 865–878.

Solomon, B., Corey-Luse, C., & Halvorsen, K. (2004). The Florida manatee and eco-tourism: Toward a safe minimum standard. *Ecological Economics, 50*(1–2), 101–115.

Stoddard, C. (1951). Wildlife economics: A neglected tool of management. *Land Economics, 27*(3), 248–249.

U.S. Department Agriculture, Farm Service Agency. (2007). *Conservation Reserve Program: Summary and enrollment statistics FY 2006.* Retrieved December 29, 2009, from http://www.fsa.usda.gov/Internet/FSA_File/06rpt.pdf

U.S. Department of Agriculture. (2008). *FY 2009 budget summary and annual performance plan.* Retrieved April 28, 2009, from http://www.obpa.usda.gov/budsum/fy09budsum.pdf

U.S. Department of Agriculture, Economic Research Service. (2009). *Conservation policy: Background.* Retrieved December 29, 2009, from http://www.ers.usda.gov/Briefing/ConservationPolicy/background.htm

van Kooten, G. C., & Schmitz, A. (1992). Preserving waterfowl habitat on the Canadian prairies: Economic incentives versus moral suasion. *American Journal of Agricultural Economics, 74*(1), 79–89.

Yandle, T. (2008). The promise and perils of building a co-management regime: An institutional assessment of New Zealand fisheries management 1999–2005. *Marine Policy, 32*(1), 132–141.

# 61

# Environmental Economics

Jennifer L. Brown

*Eastern Connecticut State University*

On the political stage, environmental issues are usually placed at odds with economic issues. This is because environmental goods, such as clean air and clean water, are commonly viewed as priceless and not subject to economic consideration. However, the relationship between economics and the environment could not be more natural.

In its purest form, economics is the study of human choice. Because of this, economics sheds light on the choices that individual consumers and producers make with respect to numerous goods, services, and activities, including choices made with respect to environmental quality. Economics is able not only to identify the reasons that individuals choose to degrade the environment beyond what is most beneficial to society, but also to assist policy makers in developing environmental policy that will provide an efficient level of environmental quality.

Because environmental economics is interdisciplinary in nature, its scope is far-reaching. Environmental economists research topics ranging from energy to biodiversity and from invasive species to climate change. However, despite the breadth of the topics covered by the community of environmental economic researchers, a reliance on sound economic principles remains the constant.

This chapter outlines the basic concepts in environmental economics, including the ways in which environmental economists might estimate the value society holds for the natural environment. Further, the corrective instruments that environmental economists can employ to correct for situations in which markets fail to achieve an efficient outcome are closely examined. This chapter also stresses the important role economic analysis plays in today's most pressing environmental issues.

## Theory

Environmental goods are those aspects of the natural environment that hold value for individuals in society. Just as consumers value a jar of peanut butter or a can of soup, consumers of environmental goods value clean air, clean water, or even peace and quiet. The trouble with these types of goods is that though they are valuable to most individuals, there is not usually a market through which someone can acquire more of an environmental good. This lack of a market makes it difficult to determine the value that environmental goods hold for society; although the market price of a jar of peanut butter or a can of soup signal the value they hold for consumers, there is no price attached to environmental goods that can provide such a signal.

To some, it may seem unethical to try to place a dollar value on the natural environment. However, there are plenty of cases in which ethics demands just that. Indeed, in cases of extreme environmental damage, such as the 1989 Exxon *Valdez* oil spill, an unwillingness to apply a value to environmental loss could be considered equivalent to stating that environmental loss represents no loss to society at all. Because of this, the assessment of appropriate damages, fines, or both, in cases such as this often depends on the careful valuation of varying aspects of the environment.

In the case of environmental policy development, insufficient evidence pertaining to the benefit that environmental

goods provide to society could easily skew the results of a cost-benefit analysis (see Chapter 27: Cost-Benefit Analysis) against environmental protection. This would, in effect, undermine the value that society holds for environmental goods and could possibly lead policy makers to believe that certain environmental regulations are not worth the costs they impose on society when, in fact, they are.

For these reasons, as well as for other reasons that are covered later in this chapter, economists have long endeavored to develop methods of accurately determining the value of environmental goods to society. This effort has led to the development of several valuation techniques.

## Valuing the Environment

### *Contingent Valuation*

Contingent valuation, or stated preferences, is a seemingly simple method of valuation that involves directly asking respondents about their values for a particular environmental good. This method is particularly useful in determining the value of environmental goods that individuals have yet to experience or may never actually experience themselves.

The Exxon *Valdez* oil spill is an example of a case in which contingent valuation provided a useful tool of valuation (Goodstein, 2008). In this case, contingent valuation was used to determine, among other things, the value that individuals place on simply knowing that a pristine Alaskan wilderness exists, even though many respondents may never actually experience this wilderness for themselves (this value is defined as *existence value*). More generally, contingent valuation methods are often used in policy development to determine the amount respondents would be willing to pay for a new, higher level of environmental quality.

However, despite its simple concept, the contingent valuation method carries with it a host of complex problems that must be taken into account for the results of a survey to be considered credible. These problems usually stem from one or more of the following: information bias, strategic bias, hypothetical bias, and starting point bias (Tietenberg, 2007). Because any type of bias can hinder the usefulness of a contingent valuation survey, special care must be taken to ensure that any bias in the answers provided by survey respondents is minimized.

With information bias, hypothetical bias, and starting point bias, respondents unintentionally misrepresent the value that they hold for an environmental good. With information bias, respondents lack enough information to form an accurate response. To avoid this type of bias, surveyors will usually provide a great deal of information to respondents pertaining to the topic of the survey.

Hypothetical bias occurs because individuals tend to respond differently to hypothetical scenarios than they do to the same scenarios in the real world. One solution to this problem is to conduct the contingent valuation surveys in a laboratory setting (Kolstad, 2000). This solution provides the surveyor with an opportunity to remind respondents to consider the financial ramifications that their responses would produce in a real-world setting. It also allows the surveyor to use experimental techniques that mimic the conditions that respondents would face in a real-world situation.

Finally, starting point bias results when respondents are influenced by the set of responses made available to them by the contingent valuation survey. The solution to this problem requires significant pretesting of a survey to ensure that its design does not influence respondents to provide biased answers (Kolstad, 2000).

Unlike the other types of response bias that can occur in a contingent valuation survey, strategic bias occurs as respondents intentionally try to manipulate the outcome of a survey. It is not always possible to eliminate intentionally biased responses. However, in general, it is best to randomly survey a large number of individuals because this will decrease the likelihood that strategic bias will undermine the overall results of the survey.

### *Revealed Preferences*

The revealed preferences method involves determining the value that consumers hold for an environmental good by observing their purchase of goods in the market that directly (or indirectly) relate to environmental quality. For example, the purchase of air fresheners, noise reducing materials, and water purification systems reveal the minimum amount individuals are willing to pay for improved air and water quality. This particular revealed preferences method is referred to as the *household production* approach.

Economists can also use revealed preferences to determine the value of clean air and clean water through differences in home prices across both pristine and polluted home locations. This particular revealed preferences method is referred to as the *hedonic approach* (Kolstad, 2000).

These approaches to valuing the environment have the advantage of relying on actual consumer choices to infer the value society holds for a particular environmental good, rather than relying on hypothetical scenarios. However, there are some environmental goods for which it can be nearly impossible to identify their value through market interactions. For example, using the revealed preferences method to determine the value that society holds for the survival of an endangered species would pose a tremendous challenge. In cases such as these, revealed preferences may not be the preferred method of valuation.

Valuation techniques are useful not only in cost-benefit analysis or in cases of extreme environmental damage but also in the more subtle cases of environmental degradation that occur as a result of market failure.

## Market Failure

As was discussed in the previous section, individual consumers will often purchase goods with an environmental component to make up for their inability to directly purchase environmental goods, thus revealing the value they hold for certain aspects of environmental quality. For

example, someone may buy a cabin on a lake in order to enjoy not only the home itself but also the pristine environment that comes with such a purchase. As long as this individual is able to exclusively incur the environmental benefits that come from owning a log cabin, the demand for log cabins will reflect the full value of both the home and the environmental goods it provides and the market for log cabins will be efficient.

Unfortunately, in the case of environmental goods, markets often fail to produce an efficient result because it is rare that any one individual can incur the full benefit (or cost) of a particular level of environmental quality. This is because environmental goods commonly suffer from the presence of externalities or a lack of property rights (see Chapter 22: Externalities and Property Rights).

There are two types of externalities: negative and positive. Negative externalities exist when individuals in society bear a portion of the cost associated with the production of a good without having any influence over the related production decisions (Baumol & Oates, 1988). For example, parents may be required to pay higher health care costs related to pollution-induced asthma among children because of an increase in industrial activity in their neighborhood.

Because producers do not consider these costs in their production decisions, they produce higher quantities of goods with negative externalities than is efficient, leading to more than the socially desirable level of environmental degradation.

As with negative externalities, positive externalities also result in inefficient market outcomes. However, goods that suffer from positive externalities provide more value to individuals in society than is taken into account by those providing these goods. An example of a positive externality can be seen in the case of college roommates sharing an off-campus apartment. Though a clean kitchen may be valued by all individuals living in the apartment, the person that decides to finally wash the dishes and scrub the kitchen floor is not fully compensated for providing value not only to himself or herself but also to the apartment as a whole. Because of this, the decision to clean the kitchen undervalues the benefits of such an action and the kitchen will go uncleaned more often than is socially desirable.

Such is the case with environmental quality. Because markets tend to undervalue goods that suffer from positive externalities, market outcomes provide a level of environmental quality that is lower than is socially desirable.

## Corrective Instruments

Once the market inefficiency relating to a particular environmental good is understood, policy makers can correct for this inefficiency by employing any number of policy instruments. Regardless of the instrument, the goal is to provide incentives to individual consumers and firms such that they will choose a more efficient level of emissions or environmental quality.

### *Command and Control*

Command and control is a type of environmental regulation that allows policy makers to specifically regulate both the amount and the process by which a firm is to reduce emissions. This form of environmental regulation is very common and allows policy makers to regulate goods where a market-based approach is either not possible or not likely to be popular. However, these regulations limit the choices that individual firms can make regarding their pollution levels. Because of this, they do not provide firms with an incentive to develop new pollution-reducing technologies (Kolstad, 2000).

### *The Coase Theorem*

Ronald Coase developed the Coase Theorem in 1960, which, although not necessarily a regulatory framework, paved the way for incentive-driven, or market-based, regulatory systems. According to the Coase Theorem, in the face of market inefficiencies resulting from externalities, private citizens (or firms) are able to negotiate a mutually beneficial, socially desirable solution as long as there are no costs associated with the negotiation process (Coase, 1960). This result is expected to hold regardless of whether the polluter has the right to pollute or the average affected bystander has a right to a clean environment.

Consider the negative externality example given previously, in which parents face soaring health care costs resulting from increased industrial activity. According to the Coase Theorem, the firm producing the pollution and the parents could negotiate a solution to this externalities issue, even without government intervention. In this example, if the legal framework in society gave the firm the right to produce pollution, the parents with sick children could possibly consider the amount they are spending on medical bills and offer a lesser sum to the firm in exchange for a reduced level of pollution. This would save the parents money (as compared with their health care costs), and the firm may find itself more than compensated for the increased costs that a reduction in emissions can bring.

If it is the parents, instead, that have a right to clean, safe air for their children (this is more typically the case), then the firm could offer the parents a sum of money in exchange for allowing a higher level of pollution in the area. As long as the sum offered is less than the cost of reducing emissions, the firm will be better off. As for the parents, if the sum of money more than compensates the health care costs they face with higher pollution levels, they may also find themselves preferring the negotiated outcome.

Unfortunately, because the fundamental assumption of the Coase Theorem (costless negotiation) often falls short, this theorem is not commonly applicable as a real-world solution. Despite this fact, the Coase Theorem is an important reminder that even in the case of complex environmental problems, there may be room for mutually beneficial compromises. This theorem also

sheds light on cases in which firms are willing to voluntarily comply with environmental regulations as well as on possible solutions to complex international environmental agreements.

*Taxation*

In 1920, Arthur C. Pigou (1920) developed a taxation method for dealing with the goods suffering from externalities. The idea behind his tax, now known as the *Pigouvian tax,* is to force producers to pay a tax equal to the external damage caused by their production decisions in order to allow the market to take into consideration the full costs associated with the taxed goods. This process is often referred to as *internalizing* an externality.

This concept can also be applied to goods that suffer from positive externalities. However, in this case, a negative tax (or subsidy) is provided to allow an individual to gain an additional benefit from providing the subsidized good. A common example of this type of subsidy can be seen each time an individual receives a tax break for purchasing an Energy Star appliance.

Of course, because the amount of the tax (or subsidy) must equal the value of the external environmental damage (or benefit) in order to correct for market inefficiencies, the valuation techniques detailed previously are crucial in the development of a sound tax policy.

*Permit Markets*

The concept of using a permit market to control pollution levels was first developed by John Dales (1968). Through this method of regulation, pollution permits are issued to firms in an industry where a reduction in emissions is desired. These permits give each firm the right to produce emissions according to the number of permits it holds. However, the total number of permits issued is limited to the amount of pollution that is allowed industry-wide. This means that some firms will not be able to pollute as much as they would like, and they will be forced to either reduce emissions or purchase permits from another firm in the industry (Barde, 2000).

Those firms able to reduce their emissions for the lowest possible cost benefit from this type of regulation. This is because these firms can sell their permits for an amount greater than or equal to the cost of their own emissions reduction, resulting in profits in the permit market. However, even firms for which it is very costly to reduce pollution experience a cost savings through this type of regulation because they are able to purchase pollution permits at a price that is less than or equal to the cost they would face if they were required to reduce emissions. Ultimately, permit markets make it less costly for an industry to comply with environmental regulations and, with the prospect of profits in the permit market, this type of regulation provides an incentive for firms to find less costly pollution reducing technologies.

## Applications and Empirical Evidence

### Valuing the Environment: Practical Applications

Both the methods of valuation and the corrective instruments described previously have been applied quite extensively to real-world environmental problems. In fact, according to Barry Field and Martha Field (2006), contingent valuation methods have been used to determine the amount respondents would be willing to pay for a myriad of environmental goods. For example, respondents have been surveyed to determine the value they would place on increased air visibility in places such as the White Mountains (located in New Hampshire) and the Grand Canyon (located in Arizona). Further, contingent valuation methods have been used to determine the value of old-growth forest preservation in the face of industrial expansion (Hagen, Vincent, & Welle, 1992).

Revealed preferences methods have increased in popularity in recent history and are commonly used by researchers to determine the value society holds for clean air and clean water. Though there are many recent cases in which researchers have used revealed preferences methods, Eban Goodstein (2008) provided a particularly useful example of the way in which this method has been used in a real-world setting.

The example given involves the decline in housing prices that occurred in the town of New Bedford, Massachusetts, in the early 1980s following severe contamination of the nearby harbor. Using the hedonic approach, economists were able to determine that those homes closest to the contamination experienced a $9000 reduction in value while the overall loss to homeowners in New Bedford was estimated to be approximately $36 million (Goodstein, 2008).

Although this type of analysis provides only a minimum value of the loss experienced due to the pollution of the harbor, it can be a valuable component in determining an appropriate fine for the firms responsible for the pollution. More generally, these results also shed light on the value that individuals place on clean water.

### Corrective Instruments: Practical Application

Though many of the concepts in environmental economics predate the 1970s, the implementation of the Clean Air Act of 1970 represents the first major application of these concepts to government policy. Through these amendments, strict ambient air quality standards were set, and in some cases, specific technologies were required for compliance (Tietenberg, 2007). This regulatory framework is consistent with the command-and-control framework described previously.

However, since the Clean Air Act Amendments of 1990, pollution taxes and permit markets have taken center stage in terms of environmental regulation. In fact, though permit markets were used in the United States as early as the 1970s, the Clean Air Act Amendments of 1990 ushered in

an era of increased popularity for this type of regulation by requiring the development of a nationwide permit market for sulfur dioxide emissions.

According to Jean-Philippe Barde (2000), the Environmental Protection Agency implemented a program in response to this requirement that was expected to result in a significant cost savings (20%–50%) as compared with other types of regulation. Further, Thomas Tietenberg (2007) asserted that the development of permit markets increased compliance with federally mandated pollution reduction requirements. (Chapter 14 of Tietenberg, 2007, provides a comprehensive overview of the effectiveness of a variety of pollution control policies.)

Additional programs have been used to reduce ozone-related emissions, including California's Regional Clean Air Incentives Market (RECLAIM), established in the Los Angeles basin, and the Ozone Transport Commission $NO_X$ Budget Program, which spans approximately 10 states in the eastern United States. (Both of these programs were originally implemented in 1994. However, the $NO_X$ Budget Program has since undergone several program modifications, including slight changes to the program name.)

The Ozone Transportation Commission program aimed to reduce nitrogen oxide emissions in participating states in both 1999 and 2003 (U.S. Environmental Protection Agency [EPA], n.d.). The results of this program, as reported by the Environmental Protection Agency, have included a reduction in sulfur dioxide emissions (as compared with 1990 levels) of over 5 million tons, a reduction in nitrogen oxide emissions (as compared with 1990 levels) of over 3 million tons, and nearly 100% program compliance (EPA, 2004).

In terms of taxation programs aimed at reducing pollution levels, Finland, Sweden, Denmark, Switzerland, France, Italy, and the United Kingdom have all made changes to their tax systems in order to reduce environmental degradation. Some of these changes include the introduction of new taxes, such as Finland's implementation of the 1990 carbon tax; other changes involve using tax revenue to increase environmental quality, such as Denmark's use of tax revenue to fund investment in energy-saving technologies (Barde, 2000).

In the United States, local grocery markets are at the center of a large tax system aimed at reducing environmental degradation: the deposit–refund system. This system effectively rewards individuals willing to return bottles and cans to an authorized recycling center. Such an incentive represents a negative tax (or subsidy) to individuals in exchange for recycling behavior that benefits society as a whole.

## Policy Implications

The policy implications of work done by environmental economists are far-reaching. As countries deal with issues such as water quality, air quality, open space, and global climate change, the methodologies developed in environmental economics are key to providing efficient, cost-effective solutions.

Although command and control remains a common form of regulation, the previous sections detail the ways that countries have begun to use market-based approaches such as taxation and permit markets within the regulatory framework.

Examples of these types of programs continue to develop. For example, in an attempt to comply with the provisions of the Kyoto Protocol, which was implemented to control greenhouse gas emissions, the European Union has established a carbon dioxide permit market aimed at reducing greenhouse gases (Keohane & Olmstead, 2007).

Even the Coase Theorem comes into play as global environmental problems demand mutually beneficial agreements to be voluntarily negotiated across countries. In fact, the Montreal Protocol, which was implemented to control emissions of ozone-depleting chemicals, makes use of a multilateral fund that compensates developing countries for the costs incurred in phasing out ozone-depleting chemicals (Field & Field, 2006). This is very similar to the example in which parents in a community may find it beneficial to compensate a polluting firm in order to induce a reduction in emissions.

## Future Directions

Because of the interdisciplinary nature of environmental economics, the discipline constantly presses forward in many directions. Many of the most pressing environmental issues involve both local and global pollutants. These range from local water quality issues to the reduction of greenhouse gas emissions.

In terms of local, regional, and national environmental issues, the application of currently available corrective instruments is quite feasible. However, an evaluation of the value of regulated environmental goods as well as the proposed regulatory instruments is still the topic of ongoing research.

In terms of global issues, such as global climate change, there is still much work to be done regarding the economic impact of changes to the earth's climate. In addition, solutions relying on government enforcement are less possible when it comes to global climate change. This means that there is likely to be more emphasis placed on voluntary compliance.

For example, in the wake of the Kyoto Protocol, there have been regional agreements that have been formed that have a reduction in greenhouse emissions as a primary goal. One such agreement, known as the Western Climate Initiative, was developed in February 2007. This initiative is a voluntary agreement between seven U.S. states and four Canadian provinces. Its goal is to reduce greenhouse gas emissions by 15% (as compared with 2005 emissions levels) by the year 2020 (Western Climate Initiative, n.d.).

Finally, countries have long suffered from the production decisions of their neighbors. However, since the availability of clean water in the border regions of developing countries remains an issue, solutions to these problems (and similar transborder problems) remain the focus of ongoing research.

## Conclusion

Environmental economics provides a set of tools that are crucial in understanding today's most pressing environmental problems. Through the use of valuation techniques such as contingent valuation and revealed preferences, economists are able to estimate the value society holds for a variety of environmental goods. These values allow policy analysts to consider the impact that a proposed public policy might have on the natural environment. Economists are also able to use these techniques to provide an accurate description of the loss that occurs in cases of both extreme environmental damage and more subtle environmental degradation that occurs daily.

Environmental economics explains the role that externalities play in excessive environmental degradation because the failure of markets to capture the full value of environmental goods consistently results in the overproduction of those goods that can damage the environment and an underprovision of those goods that improve environmental quality. Further, through corrective instruments developed by economists such as Pigou (1920), Coase (1960), and Dales (1968), environmental economics has provided society with innovative solutions to excessive environmental degradation resulting from market failure.

Finally, the application of the techniques developed by environmental economists has become increasingly popular as concern over environmental issues has become a common staple in public policy. As environmental problems continue to become increasingly complex, environmental economists continue to press forward, applying the solutions provided by the fundamentals of economics to these problems.

## References and Further Readings


Arrow, K. J., Cropper, M. L., Eads, G. C., Hahn, R. W., Lave, L. B., Noll, R. G., et al. (1996). Is there a role for benefit-cost analysis in environmental health and safety regulation? *Science, 272,* 221–222.

Barde, J.-P. (2000). Environmental policy and policy instruments. In H. Folmer & H. L. Gabel (Eds.), *Principles of environmental and resource economics: A guide for students and decision-makers* (2nd ed., pp. 157–201). Cheltenham, UK: Edward Elgar.

Baumol, W., & Oates, W. (1988). *The theory of environmental policy* (2nd ed.). New York: Cambridge University Press.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics, 3,* 1–44

Crocker, T. D. (1965). The structuring of atmospheric pollution control systems. In H. Wolozin (Ed.), *The economics of pollution* (pp. 61–86). New York: W. W. Norton.

Dales, J. H. (1968). *Pollution, property, and prices: An essay in policy-making and economics.* Toronto, Ontario, Canada: University of Toronto Press.

DeLong, J. V., Solow, R. M., Butters, G., Calfee, J., Ippolito, P., & Nisbet, R. A. (1981, March/April). Defending cost-benefit analysis: Replies to Steven Kelman. *Regulation,* 39–43.

Ellerman, A. D., Joskow, P. L., Schmalensee, R., Montero, J.-P., & Bailey, E. M. (2000). *Markets for clean air: The U.S. Acid Rain Program.* New York: Cambridge University Press.

Field, B. C., & Field, M. K. (2006). *Environmental economics: An introduction* (4th ed.). New York: McGraw-Hill.

Freeman, A. M., III. (2003). *The measurement of environmental and resource values* (2nd ed.). Washington, DC: Resources for the Future.

Goodstein, E. S. (2008). *Economics and the environment* (5th ed.). New York: John Wiley.

Hagen, D. A., Vincent, J. W., & Welle, P. G. (1992). Benefits of preserving old-growth forests and the spotted owl. *Contemporary Policy Issues, 10,* 13–26.

Keohane, N. O., & Olmstead, S. M. (2007). *Markets and the environment.* Washington, DC: Island Press.

Kelman, S. (1981, January/February). Cost-benefit analysis: An ethical critique. *Regulation,* 33–40.

Kolstad, C. D. (2000). *Environmental economics.* New York: Oxford University Press.

Kolstad, C. D., & Freeman, J. (Eds.). (2006). *Moving to markets in environmental regulation: Lessons from twenty years of experience.* New York: Oxford University Press.

Maler, K. G. (1974). *Environmental economics: A theoretical inquiry.* Baltimore: Johns Hopkins University Press.

Mitchell, R. C., & Carson, R. T. (1989). *Using surveys to value public goods: The contingent valuation method.* Washington, DC: Resources for the Future.

Pigou, A. C. (1920). *The economics of welfare.* London: Macmillan.

Sen, A. K. (1970). *Collective choice and welfare.* San Francisco: Holden-Day.

Shechter, M. (2000). Valuing the environment. In H. Folmer & H. L. Gabel (Eds.), *Principles of environmental and resource economics: A guide for students and decision-makers* (2nd ed., pp. 72–103). Cheltenham, UK: Edward Elgar.

Stavins, R. N. (2003). Experience with market based environmental policy instruments. In K. G. Maler & J. Vincent (Eds.), *Handbook of environmental economics* (Vol. 1, pp. 335–435). Amsterdam: Elsevier Science.

Stavins, R. N. (Ed.). (2005). *Environmental and natural resource economics* (6th ed.). Boston: Addison-Wesley.

Tietenberg, T. H. (1990). Economic instruments for environmental regulation. *Oxford Review of Economic Policy, 6*(1), 17–33.

Tietenberg, T. H. (1992). *Environmental and natural resource economics.* New York: HarperCollins.

Tietenberg, T. H. (2007). *Environmental economics and policy* (5th ed.). Boston: Pearson.

U.S. Environmental Protection Agency. (2004). *Acid Rain Program 2004 progress report.* Retrieved January 18, 2009, from http://www.epa.gov/airmarkets/progress/arp04.html

U.S. Environmental Protection Agency. (n.d.). *Overview of the Ozone Transportation Commission (OTC) $NO_X$ Budget Program.* Retrieved January 18, 2009, from http://www.epa.gov/airmarkets/progsregs/nox/otc-overview.html

Western Climate Initiative. (n.d.). *U.S. states, Canadian provinces announce regional cap-and-trade program to reduce greenhouse gases.* Retrieved January 18, 2009, from http://www.pewclimate.org/docUploads/Sept%2023%20PR_0.pdf

# 62

# Economics of Energy Markets

Stephen H. Karlson
*Northern Illinois University*

The improved living standards of the developed world rest on industrial processes that make intensive use of renewable and nonrenewable energy sources. Economists' attempts to understand that prosperity tend to focus on legal institutions and commercial practices rather than on resource endowments. As recently as the 1990s, David Landes's (1998) *The Wealth and Poverty of Nations* focuses on these institutions that enable people to contract with each other and to construct long-lived enterprises of large scale, rather than on resource endowments, including energy, as the foundations of that wealth. Thus, prosperity rests on division of labor, mechanization, and capital markets in a legal framework that enables people to coordinate their activities. Those elements were present in Adam Smith's (1776/1976) pin factory, which was supported by human and animal muscle power and wood. The energy source as contemporary economists understand it was not viewed as a constraint or as an opportunity in those days, although the limitations of existing sources of power imposed serious constraints on earlier economic powers.

Evolutionary biologist Matt Ridley summarizes the effect of those constraints:

> We saw a quintupling of cotton cloth output in two consecutive decades, in the 1780s and 1790s, none of it based on fossil fuels yet but based on waterpower. . . . At some point, you run out of dams. You run out of rivers in Lancashire to dam. (Bailey, 2009, pp. 50–51)

The British avoided the fate of previous economic powers by changing their power source to coal:

> By 1870 Britain is consuming the coal equivalent to 850 million human laborers. It could have done everything it did with coal with trees, with timber, but not from its own land. Timber was bound to get more expensive the more you used of it. Coal didn't get more expensive the more you used of it. (Bailey, 2009, pp. 50–51)

A similar dynamic was at work in the United States, where Alfred Chandler (1977) claims, "Coal, then, provided the source of energy that made it possible for the factory to replace the artisans, the small mill owners, and putting-out system as the basic unit of production in many American industries" (p. 77). Those transitions were not inevitable. The term *horsepower* is a marketing comparison of the early steam era, used by advocates of steam power to highlight the advantages of their machinery over those using animal power, let alone human power or waterpower, as the primary energy source. Readers will see that the quest for energy efficiency neither implies nor is implied by quests for greater economic efficiency or greater prosperity.

Energy sources have contributed to episodes of technical progress and improvements in prosperity. A recent *Economist* report (Carr, 2008) notes,

> Many past booms have been energy-fed: coal-fired steam power, oil-fired internal combustion engines, the rise of electricity, even the mass tourism of the jet era. But the past few decades have been quiet on that front. Coal has been cheap. Natural gas has been cheap. The 1970s aside, oil has been cheap. The one real novelty, nuclear power, went spectacularly off the rails. The pressure to innovate has been minimal.
>
> In the space of a couple of years, all that has changed. (p. 3)

This chapter surveys the principal features of energy markets. Energy markets, like any other markets, are environments in which prices provide incentives for substitution, conservation, and invention. Those incentives, however,

are subject to properties of markets that are the subject of advanced study, including the economics of exhaustible resources, the economics of large-scale enterprise including natural monopoly and economic regulation, the economics of common properties, and the economics of externalities. The usefulness of energy resources for industrialization and prosperity makes the resources the object of resource wars.

Daniel Yergin's *The Prize* (1991), a history of the oil business, identifies three defining influences on it. These influences—the emergence of industrial economies that use lots of energy, energy as a source of conflict, and the gains and losses from using nonrenewable energy—provide structure to his work. This chapter contains an overview of the role of energy in a modern economy, a more detailed look at the economics of exhaustible resources and of large-scale enterprises, and an examination of the public policy responses to the special problems those features pose. This chapter explores contemporary efforts to develop new sources of energy to cope with resource depletion and environmental damage. Noted also is the national security or strategic usefulness of energy supplies, without digression to the diplomatic and military consequences that are further removed from economic analysis.

## History

### Primary and Secondary Energy

The Industrial Revolution that Adam Smith and Karl Marx came to grips with is the use of machinery with some power source to augment muscle power. The term *energy* is a brief way of describing that power. The source can be what scholars and policy makers refer to as either primary or secondary energy. Primary energy refers to natural resources that can be used to provide energy, including human and animal power, wood and other combustible plants, water, wind, coal, oil, natural gas, and nuclear fission or fusion. Secondary energy refers to an energy source that requires conversion of a primary energy source. The most common form of secondary energy in the modern economy is electricity obtained from the use of fossil or nuclear fuels in a generating plant. The central steam heating plant of a university or hospital complex and the cogeneration by-product steam from a generating plant are also secondary energy sources.

### Changing Fuel Sources

The first source of primary energy other than human, animal, or waterpower was wood. In the United States, coal emerged as a domestic source around 1850. It was the single largest source of power, providing approximately 20 quadrillion Btu (quads), equal to 21 trillion megajoules, per year from 1910 through World War II. Petroleum emerged around 1900 and overtook coal in 1947. Current petroleum consumption is around 40 quads (42 trillion megajoules). Natural gas emerged as a source at about the same time and followed a similar growth curve, rising by around 1970 to 30 quads. Coal was not eclipsed by these new fuels. Current coal use is about 25 quads (U.S. Energy Information Administration [EIA], 2008, p. xx, Figure 5). The EIA expects coal use to exceed natural gas use as a primary energy source from 2015 on (Figure 6).

For 2007, the fossil fuels constitute 86.2 quads of primary energy used in the United States, comprising 22.8 quads from coal, 23.6 quads from natural gas, and 39.8 quads from petroleum. An additional 8.4 quads originate as nuclear electric power, with 6.8 quads provided by what the EIA (2008) classifies as renewable energy, aggregating "conventional hydroelectric power, biomass, geothermal, solar/photovoltaic, and wind." Annual energy consumption was 101.6 quads, of which industrial users consumed 32.3 quads, transportation 29.1, commercial enterprises 18.4, and residences 21.8 (EIA, p. 3, Diagram 1).

### Energy Intensity Diminishes

Although industrialization means an increase in an economy's use of energy, the intensity with which an economy uses its energy tends to diminish. For instance, Natural Resources Canada (2009) reports that the aggregate energy intensity of Canadian industry was 13,000 Btu (13.6 megajoules) per 1997 dollar of gross domestic product in 1990, which falls to 11,000 Btu (11.3 megajoules) per 1997 dollar in 2006.

Declining energy intensity of industries and of entire economies is characteristic of most industries in most countries, although readers will see that energy prices or attempts to achieve greater energy efficiency are not necessarily driving these changes.

Na Liu and B. W. Ang (2007) evaluate the research on energy intensity in a paper that, although not explicitly a survey, includes references to numerous other surveys of the evidence. The most common outcome researchers identify is reduced energy intensity over time, accounted for either by industry-wide changes in energy intensity, which they refer to as intensity effects, or by changes in the mix of products within an industry, which can be a shift to a more or a less energy-intensive mix of products, which they refer to as structural effects. In Liu and Ang's review, the modal outcome is industry-wide reductions in energy intensity and a less energy-intensive mix of products, although many studies uncover a switch to a more energy-intensive mix of products that is offset by industry-wide reductions in energy intensity. The article focuses on improving index number techniques, leaving work on the sources of changing intensity, whether driven by prices or by factor-neutral technical change, for future research. Readers will see that such research will be of value for

policy makers coping with resource depletion and environmental degradation as consequences of energy use.

## Economic Theories for Modeling Energy Markets

Although energy markets offer economists ample opportunity to apply the traditional tools of supply and demand and those tools have been used to great effect in changing public policies, there are some features of energy markets where special models provide additional insight. Many primary energy sources are depletable, providing opportunities to use the theory of exhaustible resources. Many primary and secondary energy producers are large enterprises, where the theory of a natural monopoly is useful. Because energy markets sometimes involve competition of a few firms, for use of a common source or with pollution of a common sink, game-theoretic approaches (generally based on the prisoners' dilemma) help structure thinking about public policy. The large scale of energy enterprises is often a consequence of special equipment such as refinery vessels or turbogenerators, and energy users often make investments in special equipment, which in the home might be a refrigerator or an air conditioner or in the factory might be an energy management system or a new steel furnace. The common feature of all such special equipment is that it is lumpy (one can speak of a smaller refrigerator but not of half a refrigerator) and irreversible (that steel furnace cannot be turned into a slow cooker, and the resale market for steel furnaces is thin).

### Exhaustible Resources

Energy consumers have recognized for some time that their primary energy sources can be depleted. Deforestation inspired an early generation of conservationists in the mid-nineteenth century. The current generation of environmentalists might be surprised to learn that fears of the extinction of the whales arose at about the same time (Yergin, 1991). Where capital markets and property rights exist, there are incentives for people to invest in replacing the stock that they take. Tomislav Vukina, Christiana Hilmer, and Dean Lueck (2001) study the price of Christmas trees in North Carolina using a model of those incentives to provide the hypotheses they test. The same model can be used to consider the replanting of trees, sugarcane, or corn to provide feedstock for biomass fuels. There is no generalization of the theory to the oceans, which is regrettable for the whales.

One of the changes Geoffrey Carr (2008) refers to in contemporary energy markets is the dawning fear that the oil reserves will be depleted. In popular parlance, the expression *peak oil,* coined by Shell Oil petroleum geologist M. King Hubbert, refers to that time at which more than half the world's proven reserves have been used or to the time at which the cost of extracting the oil reserves begins to rise. One could speak of peak coal in the same way: That neither Ridley nor Chandler characterized coal as subject, with increased use, to rising prices reflects improvements in the technology for mining coal as well as price competition from other sources of primary energy, including oil and natural gas.

The theory of valuing an exhaustible resource provides both a logical structure to the peak oil problem and an explanation of the nondepletion of coal. The theory begins with Harold Hotelling's (1931) model. Although the mathematics (calculus of variations) proved daunting to economists of the day, the general principle is simple. A stock of an exhaustible resource is a capital asset. A wealth-maximizing holder of such an asset will use it in such a way as to be indifferent about the choice between consuming it now and holding it for later use. That indifference principle suggests the price of the resource will increase at the rate of interest, if the owner is in a competitive market. If the owner is a monopolist, the marginal revenue increases at the rate of interest. If extraction costs, either constant or contingent on the rate of depletion, are present, the argument becomes more complicated, but the general principle still applies. The results change in the presence of a backstop technology, which will replace the resource before it is depleted. Some models of exhaustible resources treat the time that a backstop technology becomes available as predetermined. The complementary problem, in which the Hotelling principle provides an incentive to develop the backstop technology, has not been investigated as intensively. Christopher Harris and John Vickers (1995) suggest a promising approach for such investigation.

Shantayanan Devarajan and Anthony Fisher (1981) revisit Hotelling's paper, identifying further improvements on the model and providing empirical extensions. Robert Pindyck (1980) proposes a number of extensions, based on uncertainty in the resource markets. Empirical tests are difficult, owing to difficulties obtaining the price at which the exhaustible resource itself trades, because the resource itself is often extracted by companies that transform it into some other product before selling it. That is true of vertically integrated oil companies, which is one reason the model has not been used to test the peak oil hypothesis empirically, although James Hamilton (2008) addresses peak oil in light of the Hotelling principle, and C.-Y. Cynthia Lin, Haoying Meng, Tsz Yan Ngai, Valeria Oscherov, and Yan Hong Zhu (2009) offer a theoretical explanation for what they characterize as trendless oil prices where technical progress is a possible response to rising energy prices. Readers will also see that the absence of a trend in the price of oil or coal reduces the incentive to develop a backstop technology.

### Large-Scale Enterprise

A second component of the theory of energy markets is that of markets that cannot be described using the perfectly

competitive model. The first large-scale enterprises of the Industrial Revolution included the coal mines, along with the canals and later the railroads that came into being to transport the coal. From the beginning, public policy makers had to choose whether to create public enterprises or to rely on private investment. The roads and canals were among the first public internal improvements of the United States and Britain. The railroads and the mines tended to be private enterprises at first, although that is not universally true. Public policy makers had to improvise new legal structures both to make possible and to restrain those enterprises. In energy markets, the institutions of antitrust, regulation, and public enterprise each play a role.

*Antitrust*

The coal economy tended to involve smaller, less vertically integrated firms in which the mining, transportation, and retailing were functions for distinct businesses. The rudimentary rules of contract and liability, perhaps supported by the inchoate intellectual basis there was for understanding a competitive economy, sufficed as an energy policy. That was not the case for the oil economy, in which the Standard Oil Trust became an early example of a vertically integrated firm, combining refining with pipeline transportation and retail distribution, as well as working with railroads to obtain more favorable prices for transporting its inputs and outputs than its smaller competitors could bargain for. The trust emerged as a producer and distributor of kerosene for lighting and cooking well before the diffusion of private automobiles, a development that offered the oil companies the opportunity to vertically integrate into operating service stations.

Yergin's (1991) *The Prize* provides a thorough overview of the emergence of the large oil companies. The worldwide extraction of crude oil, its refinement into fuels, lubricants, and petrochemicals, and its distribution to consumers led to firms of great scale and scope. In this expansion, entrepreneurs had many opportunities to get rich. The size of the stakes provided ample opportunities for corporations and governments to engage in acts of corruption and for owners to work together to take advantage of consumers.

Although the *Standard Oil Company of New Jersey v. United States* (1911) decision, which antitrust scholars call significant for its enunciation of the rule of reason as a general principle for enforcing United States antitrust laws, did not become the landmark case out of any special desire to make an example of the Standard Oil Trust, the legend of grasping and predatory oil barons the case inspires lives on to this day. John McGee (1958) interprets the evidence in the *Standard Oil* case to suggest that the oil trust, although clearly intending to monopolize sales of petroleum products, did so in such a way as to raise, rather than lower, its profits. He finds no evidence of the company selling at a loss to eliminate rivals.

The barons, however, more frequently operate in concert, rather than as a monopoly firm. In part, cooperative behavior is the only option in an industry where the dominant firm has been broken up by antitrust action. Cooperation, however, is a logical outcome for competing firms making common use of oil fields or pipelines. A similar logic applies to natural gas producers, and electricity producers share a common electricity transmission grid. Where firms have strong incentives to cooperate, government's best response might be to supervise the competition, or it might be to operate the energy companies itself. The Organization of Petroleum Exporting Countries turns out to be an attempt by multiple governments to cooperate in such supervision.

*Regulation*

Establishing ownership of crude oil is not easy. An oil pool might extend under several parcels of property or under a national boundary. Common-law methods such as the rule of capture, which works for a hunter taking an animal on land, provide incentives for each oil producer to extract oil from under its property more rapidly, before somebody else pumps it out (Yergin, 1991). That rule also provides incentives for producers to engage in slant drilling, where the wellhead is on the producer's land but the well draws from oil under a neighbor's land. The consequence is uneconomic extraction of the oil, because additional wells dissipate the natural gas that provides pressure to push the oil to the surface. Some oil that would otherwise be extracted remains in the ground, and producers invest in pumps that they would otherwise not have to install. When the competition is between nations, drilling under national boundaries becomes a casus belli, as it did most recently in Iraq's 1990 invasion of Kuwait.

The Texas Railroad Commission pioneered the use of an independent regulatory commission to manage the output of an oil field. Under a policy known as prorationing, each existing well received a production quota worked out with the intent of obtaining the maximum economic value of the field but often with the effect of enriching the firms subject to that regulation (Wilcox, 1971).

A prorationing policy with well-informed regulators can match the resource extraction behavior of competitive producers drawing down the resource in a Hotelling-optimal way. Those regulators can also match the resource extraction behavior of a monopoly, in which the marginal revenue from extraction rises at the rate of interest. The two outcomes are of more than academic interest, because the common property problem (Hardin, 1968) and the cartel problem (Osborne, 1976) can both be interpreted as prisoners' dilemmas, in which the regulator can prohibit the individually rational but collectively suboptimal dominant strategy equilibrium, which is too rapid a depletion of the common property in the former but means lower prices for consumers in the latter. Yergin (1991) credits the Texas

Railroad Commission with providing the Organization of Petroleum Exporting Countries, commonly called a cartel, with a model for their production quotas.

The large-scale enterprise provides a second, different rationale for government regulation when the enterprise is sufficiently large relative to its market that it is a natural monopoly. Perhaps the natural monopoly arises because the duplication of facilities, such as electric or gas distribution lines in a community, implies investments whose costs exceed any benefits that consumers might get from competitive supply of the electricity or the gas. Or perhaps the infrastructure involves large sunk costs, such that competing firms engage in Bertrand price competition down to avoidable incremental costs, risking the long-term profitability of both companies. In economic theory, a natural monopoly is an industry in which the cost function is subadditive, meaning any division of the outputs among two or more firms involves higher costs than a single firm would incur, over outputs likely to be observed in the industry's market (Baumol, Panzar, & Willig, 1988). Where a single firm can serve the market more cheaply than two or more firms, that firm might be able to price like a profit-maximizing monopolist, with the attendant allocative inefficiencies.

The public utility concept, in which a company obtains a legal monopoly subject to supervision by an independent regulatory commission, emerged as a more flexible replacement for legislative or court supervision or for a public ownership that some policy makers viewed as ideologically suspect and others saw as subject to corruption. Before the Great Depression, most states had regulatory commissions, and by 1966, they had diffused to all the states (Phillips, 1969). As government agencies, however, regulatory commissions can be subject to the same public choice dynamics that confront public enterprises or governments themselves and make them imperfect instruments of control (Hilton, 1972).

Public ownership of the natural monopoly offers an alternative to direct regulation. Economic theory argues that natural monopoly is not a sufficient condition for monopolistic exploitation of consumers (Baumol et al., 1988, particularly chap. 8). These alternatives receive consideration in subsequent sections of the article.

### *Deregulation*

Changes in the direct regulation of energy companies reflected both failures of the regulatory apparatus and improvements of market institutions. Alfred Kahn (1988, pp. xv–xvii) offers a useful summary of the events. Put briefly, regulatory failures in energy markets provided economists and policy makers with incentives to consider alternatives to direct regulation.

In natural gas, regulators had the responsibility of determining wellhead prices for natural gas, interstate transmission rates for common-carrier pipelines facing increasing returns to scale, and local delivery rates for consumers served by municipal gas mains that operated under canonical natural monopoly conditions. The outcome, however, was what Paul MacAvoy (1971) calls a "regulation-induced shortage" of natural gas. Because natural gas fields consist of multiple independently operated wells producing natural gas jointly with crude oil, standard formulas to price the output well by well or field by field broke down in administrative complexity. Today's regulatory structure, in which city distribution companies and interstate transmission companies (where natural monopoly arguments make some sense) remain regulated while the gas wells enjoy relative freedom to compete in price, emerged as a less cumbersome alternative.

In electricity, a combination of perceived difficulties in regulating the enterprises with improvements in the implementation of market pricing led to the partial or full deregulation of the electric utilities. George Stigler and Claire Friedland (1962) suggest that regulators had relatively little effect on the price of electricity because electric utilities had relatively little monopoly power. The cost and duration of regulatory cases led regulators to circumvent their own procedures with automatic fuel adjustment clauses that raised production costs (Gollop & Karlson, 1978). Subsequent empirical research could not reject the hypothesis that regulated electric utilities were charging monopolistic prices (Karlson, 1986). At the same time, theoretical work considered the possibility of competition for the right to operate a natural monopoly, which in the work of Harold Demsetz (1968) takes the form of an auction, with the bidder offering to operate the service for the lowest price obtained the franchise, and in the work of Baumol et al. (1988) and extensive follow-on research takes the form of sufficient conditions under which potential competition compels a natural monopoly to price efficiently.

Paul Joskow (1997) provides an overview of the changed circumstances leading to deregulation. The transition to a deregulated environment came with difficulties, the most famous of which is the California power crisis of early 2001. Severin Borenstein (2002) evaluates what went wrong and suggests some directions for future improvements of policy.

### *Public Enterprise*

In much of the world, the energy industries are private enterprises simply as a matter of course, with no ideological statement intended by the government or understood by the citizens (Viscusi, Harrington, & Vernon, 2005, chap. 14; Wilcox, 1971, chap. 21). In the United States, nuclear electricity is a by-product of research into the use of nuclear fission for purposes other than weapons. Contemporary efforts to develop solar, wind, and biofuels involve federal subsidies and public–private partnerships. The effectiveness

of many of these projects will provide term paper topics for students later in the century.

### Irreversible Investments

Many investments involve irreversible commitments to purchase equipment that cannot be easily converted to other uses. Electricity generating plants are large examples of such investments. Home air conditioners and refrigerators are smaller examples. All three are technologies that have the potential to reduce the economy's energy intensity as new units replace older ones, in the first case by reducing the energy intensity of the generating system, in the second and third by reducing household energy use. All three have been the subject of economic research suggesting that investors hold out for returns on their investment that exceed the opportunity cost of capital.

That reluctance, in seeming defiance of all models of rational investment behavior, was observed so frequently among energy producers and energy users that it received a special name, the *energy paradox.* Kenneth Train (1985) offers an early survey of research that seems to identify a reluctance to invest. Kevin Hassett and Gilbert Metcalf (1993) suggest that investors in energy conservation technologies require a rate of return of about four times the cost of capital for the investments they make. In subsequent research, the same authors evaluate the quality of the data used in consumer studies to suggest that "the case for the energy paradox is weaker than has previously been believed" (Metcalf & Hassett, 1999, p. 516).

That reluctance to invest is neither necessarily suboptimal nor necessarily paradoxical. The act of making an irreversible investment involves the exercise of a real option—namely, to defer the investment until economic circumstances are more favorable, where favorable can mean a higher price for the electricity the generating capacity produces—or a higher price for the electricity used to power the refrigerator or air conditioner. Investor behavior is thus a manifestation of economic hysteresis. The general theory of irreversible investments has been the subject of extensive analysis, much of it specifically inspired by observed behavior in energy markets. Avinash Dixit's (1992) "Investment and Hysteresis" is a straightforward introduction to the topic. Dixit and Pindyck's (1994) *Investment Under Uncertainty* provides comprehensive treatment of several different models of irreversible investment, with energy applications that will reward careful study.

The recent (late-2007 to mid-2009) swings in crude oil prices call for such research. A permanently higher price, or a price rising at or with the interest rate, is a stronger incentive to work on a replacement technology. A falling price weakens that incentive. In the irreversible investment models, greater price volatility also weakens the incentive. Where there is currently no backstop technology available, delayed invention has the potential to leave an economy with a depleting resource and no replacement in development.

## Energy and the Environment

The interaction between energy uses and environmental consequences presents researchers and policy makers with substantial challenges. On one hand, worldwide economic development means improved living standards for people whose parents or grandparents might have lived their entire lives in extreme poverty. On the other hand, that development involves additional demands for the stocks of nonrenewable resources, additional pollution from the use of the carbon-based and nuclear primary energy sources, and additional pressures on water and land that has uses other than as energy sources.

That global development is substantial. Thomas Friedman (2008) uses the expression *Americum* to refer to "any group of 350 million people with a per capita income above [U.S.] $15,000 and a growing penchant for consumerism" (p. 50). That figure once described to two populations, primarily in North America and western Europe. There are at least two more today, one each in India and in China, and an environmental consultant Friedman cites expects a world of eight or nine Americums, which Friedman characterizes as "America's carbon copies." That's an ironic expression, referring to the potentially adverse economic and environmental consequences of that development.

There are several trade-offs at work. First, reductions in the energy intensity of the world economy are not necessarily improvements in the efficiency or the prosperity of the world economy. One does not have to contemplate a return to the less energy-intensive world economy of 1700, when living conditions were worse for everyone. The economic model of substitution in production provides the explanation for today's economy: Allocative efficiency is the equating of marginal products scaled for input prices. A firm that reduces its energy use irrespective of the opportunity cost of other inputs, or a public policy that mandates reductions in energy use without regard to those opportunity costs, reduces output and the allocative efficiency of the economy. Adam Jaffe, Richard Newell, and Robert Stavins (1999) describe several differing visions of lowered energy intensity. Three are relevant to this chapter. First, there is a technologist's optimum, in which economic efficiency is irrelevant as long as energy efficiency is increased. From an economic perspective, that outcome ignores the opportunity costs of the inputs that produce the outputs from which the derived demand for energy arises. Second, there is a theoretical social optimum, in which the cost of implementing energy-efficiency policies is irrelevant but the opportunity cost of other inputs is relevant. From an economic perspective, that outcome abstracts from the frictions of developing and implementing public policy. Third, they suggest a true social optimum, comprising those corrective policies that

pass a cost-benefit test. The paper includes a useful list of references that supplements those noted in this chapter. The authors suggest that "market signals are effective for advancing [diffusion]" of new technologies, but imposition of minimum standards for energy efficiency, such as automotive fuel economy requirements, may not be. The market signals, however, can be incentives to adopt a technology because of its potential to lower the adopting firm's costs (a process technical change) rather than because of the energy savings it promises (a price-induced technical change). For example, Gale Boyd and Stephen Karlson (1993) suggest that the process incentive, rather than the price incentive, induced steel companies to install new steel-making technologies.

The achievement of any form of energy efficiency is more difficult because energy use produces negative nonpecuniary externalities. Completing energy markets, by taxation or regulation to address those externalities, is therefore likely to be a long-lived project for researchers and for policy makers. Such market completion might foster development of replacements for nonrenewable energy resources. It also changes the incentives energy consumers will face. The business that seeks all profitable opportunities to conserve on fuel use, for instance, currently faces prices that do not reflect the mortality or morbidity of a smoggy city or of proximity to a uranium mine or a nuclear waste pile, let alone the potential lost output that would follow a melting of the polar ice caps. The cost-benefit test that yields the true social optimum of Jaffe et al. (1999) requires somebody, or some collectivity, to determine what benefits and costs make up that test.

More recent research contemplates policy mixes to achieve compliance with tighter environmental standards, such as the Kyoto Protocol targets, while reducing the economic welfare or efficiency losses that compliance might imply. William Pizer (2002) simulates several policy changes looking forward to 2010. He suggests that policy makers combine mitigation policies, rather than rely on emissions targets or corrective taxes alone, to achieve greater efficiency gains. Bob van der Zwaan, Reyer Gerlagh, Ger Klaassen, and Leo Schrattenholzer (2002) introduce endogenous technical change, perhaps induced by environmental policies, into several macroeconomic simulations to suggest that improvements in technologies other than fossil-fuel-using technologies are more promising at reducing carbon emissions. The results of these simulations are not surprising, although they suggest opportunities for research on the actual evolution of new energy sources and new energy-conserving technologies in 2010 and beyond, perhaps in combination with work on irreversible investments and improved trading regimes for pollution permits.

Second, efforts to mitigate climate change without returning world standards of living to those of 1700 involve additional equity and efficiency trade-offs. Mitigation in an equitable way poses problems that may not be the comparative advantage of economists. Richard Tol (2001) summarizes the challenge: The poorer countries, as measured by their low energy use, also face the greater harm from climate change. "Greenhouse gas emissions and vulnerability to climate change show a strong negative correlation. This is the moral issue at the heart of the climate problem" (p. 71). He describes his paper as "academic constructs" with the potential to "help to inform further thinking about how to handle the enhanced greenhouse effect" (p. 84).

Third, although renewable energy sources provide a way around depletion of the nonrenewable sources, those sources also involve trade-offs. Fuels that make use of biomass, including ethanol and vegetable oils, are carbon compounds. The act of growing the plants can serve as a carbon sink, but the net carbon balance need not be positive. Waterpower cannot escape the running out of rivers to dam. Reservoirs pose a common property problem in which maintaining sufficient depth for electricity or other industrial use means holding water back from downstream drinkers or recreational users. Wind power requires a connection to the electric transmission grid. The most reliable winds are in sparsely settled parts of the United States, and the power grids are where the people are. Thomas Ackermann (2005) provides a comprehensive survey of the technical challenges facing wind-power producers. Finally, land occupied by solar collectors is sometimes not available for other uses. Each of these technologies further involves an irreversible investment facing competition from exhaustible resources whose prices neither follow Hotelling paths nor incorporate the effects of negative externalities.

## Conclusion

Energy markets allocate the primary and secondary energy sources that have relaxed the constraints of human and animal power on creating, producing, and exchanging. Those markets have also called for economic analysis using models other than the standard perfectly competitive model. Those models have suggested public policy reforms and reforms to those public policies. The use of primary energy requires that producers, consumers, and policy makers deal with resource depletion and environmental degradation. The challenges of these problems will continue to provide economists with theoretical and empirical research opportunities.

## References and Further Readings

Ackermann, T. (Ed.). (2005). *Wind power in power systems.* Chichester, UK: John Wiley.

Bailey, R. (2009). Chiefs, thieves, and priests. *Reason, 40*(9), 49–54.

Baumol, W. J., Panzar, J. C., & Willig, R. D. (1988). *Contestable markets and the theory of industry structure* (rev. ed.). San Diego, CA: Harcourt Brace Jovanovich.

Borenstein, S. (2002). The trouble with electricity markets: Understanding California's restructuring disaster. *Journal of Economic Perspectives, 16*(1), 191–211.

Boyd, G. A., & Karlson, S. H. (1993). The impact of energy prices on technology choice in the United States steel industry. *Energy Journal, 14*(2), 47–56.

Carr, G. (2008, June). The power and the glory. *Economist, 387,* 8585.

Chandler, A. D., Jr. (1977). *The visible hand: The managerial revolution in American business.* Cambridge, MA: Harvard University Press.

Demsetz, H. (1968). Why regulate utilities? *Journal of Law and Economics, 11*(1), 55–65.

Devarajan, S., & Fisher, A. C. (1981). Hotelling's "economics of exhaustible resources": Fifty years later. *Journal of Economic Literature, 19*(1), 65–73.

Dixit, A. K. (1992). Investment and hysteresis. *Journal of Economic Perspectives, 6,* 107–132.

Dixit, A. K., & Pindyck, R. S. (1994). *Investment under uncertainty.* Princeton, NJ: Princeton University Press.

Friedman, T. L. (2008). *Hot, flat, and crowded: Why we need a green revolution and how it can renew America.* New York: Farrar, Straus and Giroux.

Gollop, F. M., & Karlson, S. H. (1978). The impact of the fuel adjustment mechanism on economic efficiency. *Review of Economics and Statistics, 60*(4), 574–584.

Hamilton, J. D. (2008). *Understanding crude oil prices* (UCEI Energy Policy and Economics Working Paper No. 023). Retrieved January 29, 2009, from http://www.ucei.berkeley.edu/PDF/EPE_023.pdf

Hardin, G. (1968). The tragedy of the commons. *Science, 162*(3859), 1243–1248.

Harris, C., & Vickers, J. (1995). Innovation and natural resources: A dynamic game with uncertainty. *RAND Journal of Economics, 26*(3), 418–430.

Hassett, K. A., & Metcalf, G. (1993). Energy conservation investment: Do consumers discount the future correctly? *Energy Policy, 21,* 710–716.

Hilton, G. R. (1972). The basic behavior of regulatory commissions. *American Economic Review Proceedings, 62*(2), 47–54.

Hotelling, H. (1931). The economics of exhaustible resources. *Journal of Political Economy, 39*(2), 137–175.

Jaffe, A. B., Newell, R. G., & Stavins, R. N. (1999). *Energy-efficient technologies and climate change policies: Issues and evidence* (Resources for the Future Climate Issue Brief No. 19). Washington, DC: Resources for the Future.

Joskow, P. A. (1997). Restructuring, competition, and regulatory reform in the U.S. electricity sector. *Journal of Economic Perspectives, 11*(3), 119–139.

Kahn, A. E. (1988). *The economics of regulation: Principles and institutions.* Cambridge: MIT Press.

Karlson, S. H. (1986). Multiple-output production and pricing in electric utilities. *Southern Economic Journal, 53,* 73–86.

Landes, D. S. (1998). *The wealth and poverty of nations: Why some are so rich and some so poor.* New York: W. W.Norton.

Lin, C.-Y. C., Meng, H., Ngai, T. Y., Oscherov, V., & Zhu, Y. H. (2009). Hotelling revisited: Oil prices and endogenous technical progress. *Natural Resources Research, 18*(1), 29–38.

Liu, N., & Ang, B. W. (2007). Factors shaping aggregate energy intensity trend for industry: Energy intensity versus product mix. *Energy Economics, 29,* 609–635.

MacAvoy, P. A. (1971). The regulation-induced shortage of natural gas. *Journal of Law and Economics, 14,* 167–199.

McGee, J. S. (1958). Predatory price cutting: The Standard Oil (N. J.) case. *Journal of Law and Economics, 1,* 137–169.

Metcalf, G. E., & Hassett, K. A. (1999). Measuring the energy savings from home improvement investments: Evidence from monthly billing data. *Review of Economics and Statistics, 81,* 516–528.

Natural Resources Canada, Office of Energy Efficiency. (2009, January). *Industrial energy intensity by industry.* Retrieved January 25, 2009, from http://oee.nrcan-rncan.gc.ca/corporate/statistics/neud/dpa/tableshandbook2/agg_00_6_e_1.cfm?attr=0

Osborne, D. K. (1976). Cartel problems. *American Economic Review, 66*(5), 835–844.

Phillips, C. F., Jr. (1969). *The economics of regulation* (Rev. ed.). Homewood, IL: Richard D. Irwin.

Pindyck, R. S. (1980). Uncertainty and exhaustible resource markets. *Journal of Political Economy, 88*(6), 1203–1225.

Pizer, W. A. (2002). Combining price and quantity controls to mitigate global climate change. *Journal of Public Economics, 85,* 409–434.

Posner, R. A. (1969). Natural monopoly and its regulation. *Stanford Law Review, 21*(3), 548–643.

Smith, A. (1976). *An inquiry into the nature and causes of the wealth of nations.* Chicago: University of Chicago Press. (Original work published 1776)

Stigler, G. J., & Friedland, C. (1962). What can regulators regulate? The case of electricity. *Journal of Law and Economics, 5,* 1–16.

Tol, R. S. J. (2001). Equitable cost-benefit analysis of climate change policies. *Ecological Economics, 36,* 71–85.

Train, K. (1985). Discount rates in consumers' energy-related decisions: A review of the literature. *Energy, 10,* 1243–1253.

U.S. Energy Information Administration. (2008, April). *Annual energy review 2007.* Washington, DC: Department of Energy. Retrieved January 24, 2009, from http://www.eia.doe.gov/emeu/aer/contents.html

van der Zwaan, B. C. C., Gerlagh, R., Klaassen, G., & Schrattenholzer, L. (2002). Endogenous technical change in climate change modelling. *Energy Economics, 24,* 1–19.

Viscusi, W. K., Harrington, J. E., Jr., & Vernon, J. M. (2005). *Economics of regulation and antitrust* (4th ed.). Cambridge: MIT Press.

Vukina, T., Hilmer, C. E., & Lueck, D. (2001). A Hotelling–Faustmann explanation of the structure of Christmas tree prices. *American Journal of Agricultural Economics, 83*(3), 513–525.

Wilcox, C. (1971). *Public policies toward business.* Homewood, IL: Richard D. Irwin.

Yergin, D. (1991). *The prize: The epic quest for oil, money, and power.* New York: Simon & Schuster.

# 63

# Political Economy of Oil

Yahya M. Madra
*Gettysburg College*

In contrast to more narrow economic analyses of oil, where the focus lies on the price formation and the conditions under which the "right" price for this nonrenewable resource can be obtained, political economy of oil studies the conditions and consequences of the production, appropriation, distribution, and consumption of oil (and oil-related products) by taking into account social relations of power (at local, regional, and global scales), cultural codes of consumption, institutional structures of surplus extraction, and ecological impacts of human activity. In this sense, again in contrast to economics of oil, political economy of oil is decidedly interdisciplinary—moving beyond economics, it draws upon political science and international politics, sociology and cultural studies, geology and geography, and ecology. In what follows, the contours of a political economy approach (as distinct from the standard neoclassical approach) to oil will be outlined. Beginning with the status of oil as a nonrenewable resource, the chapter will discuss the nature of the global demand for oil and will offer a historical account of the concrete socioeconomic processes (including the processes of price making, market maintenance, and state formation) within which the price of oil is determined. In the process, a general political economy framework will be developed to explain why the increases in the price of oil may not result in the changes in production technologies and consumption patterns that are necessary to move beyond petroleum. In short, the chapter will argue that not only producer but also consumer petro-states suffer from the so-called oil curse.

## Oil as a Nonrenewable Resource: Malthusianism and Its Limits

Oil is a fossil fuel, 150- to 300-million-year-old "solar energy" buried underground. Up until the Industrial Revolution, most of the world's energy was supplied from renewable sources. But with the Industrial Revolution and the depletion of wood in England in the early eighteenth century, the transition to technologies that run on nonrenewable sources of energy occurred as production and transportation began to rely increasingly on coal (Mitchell, 2009). Earlier in the twentieth century, we observe another shift to a new nonrenewable resource pioneered first by the introduction of battleships that run on oil to the Royal Navy and then by the mass production of cars in the 1920s (Yeomans, 2004; Yergin, 1991). Today, our highly industrialized, and predominantly capitalist, world economy continues to be heavily dependent on this nonrenewable energy source.

This state of dependence inevitably begs the question of the exhaustion of this nonrenewable resource. How far is it? How much oil is in the ground? What is the rate at which new oil reserves are found? How difficult is to extract them? How reliable is the reserve-to-production ratio that indicates the length of time the remaining oil reserves will last if the production will continue at the current levels? What is the rate at which the consumption of oil is growing? Even though the answers to these questions are heavily contingent upon the assumptions one is willing to make in calculations, if one were to accept the "optimistic" predictions of the International Energy Agency (from the 2004 edition of their *World Energy Outlook*) regarding the possibility of reaching peak production sometime between 2013 to 2037, it will be probably safe to assume that the current reserve-to-production ratio of 42 years (reported in the 2009 edition of British Petroleum's *Statistical Review of World Energy*) will not improve drastically in the coming decades.

Nevertheless, the picture is not that simple. For instance, it is important to recognize that what makes the geological approach to peak oil so powerful in the minds of the general public and research community is Hubbert's

(1956) success in predicting the U.S. (a limited territory) peak in 1970. Yet, in order to extend the argument to the world scale and to claim that the future production rate depends linearly on the unproduced fraction (Deffeyes, 2006, p. 40), a number of assumptions must be made. For instance, to assume that the aggregation of all regional annual oil production distribution fits a bell curve is to assume that the future path of global oil production will inversely mirror its past trajectory where a cheap and expanding oil supply will be followed beyond the peak by an expensive and decreasing one (Caffentzis, 2008, p. 315). One underlying assumption is that the larger and easily accessible fields will be depleted earlier than smaller and more difficult to extract fields (e.g., Klare, 2008a, p. 42). Nevertheless, for this to be true, one also has to assume the existence of a fully competitive market where the price of crude oil net of extraction costs grows steadily at a rate equal to the rate of interest (Hotelling, 1931). Yet, if one were to make these assumptions, as it will be shown below, there would be no reason to worry about oil depletion: The world economy, in response to price increases, will change its production technology and consumption patterns gradually, moving away from oil and instead substituting alternative, less scarce, resources.

Nevertheless, Hubbert's (1956) calculations did not take price into account (Deffeyes, 2006, p. 41). But more important, even if one were to take the price of oil into account, one would have to do so by acknowledging that historically, it has never been determined in competitive markets and that it has not been growing along an efficient extraction path. On the contrary, the long-run secular path of the price of oil reflects either the fall in costs brought about by technological progress or the various historical transformations and shifts in the market structure and the geopolitics of global oil production (Roncaglia, 2003). Moreover, the relatively high level of prices since the 1970s has allowed for the exploitation of expensive and difficult to extract oil fields (e.g., North Sea), whereas large and easy to extract fields continue to remain relatively underexploited (Roncaglia, 2003, p. 646f). And finally, in the past three decades, in part due to more effective exploration technologies and in part due to the high price of oil, the oil reserves have increased steadily, bringing the reserve-to-production ratio from 29 years in 1980 to 42 years in 2008 (British Petroleum, 2009).

Without doubt, the growing body of research and popular literature on "peak oil" has generated an increasing literacy among the general public about the coming end of the oil era. Nevertheless, the effects of this literature on the public perception have so far been mixed. On one hand, the dissemination of this particular kind of geological "petro-knowledge" increased the social legitimacy of high energy prices in the eyes of the general public. It is worthwhile to note that international oil companies, in 1971, as soon as it became clear that the formation of the Organization of the Petroleum Exporting Countries (OPEC) and the nationalization of oil in the Middle East would limit their control over the flow of oil and lead to increases in the price of crude oil, "abruptly abandoned their cornucopian calculations of oil as an almost limitless resource and began to forecast the end of oil" (Bowden, 1985; cf. Mitchell, 2009, p. 419). On the other hand, the often apocalyptic and Malthusian tone of the books and documentary films in the genre make it a daunting task for ordinary citizens to tackle this complex issue. In short, while simple geological analyses of "peak oil" provide a necessary baseline in beginning to think about the political economy of oil, they fail to do justice to the complexity of the political, economic, and cultural processes that shape the exploration, production, distribution, and consumption of oil.

## Oil as a Means of Consumption and Exploitation: Global Oil Demand

President Franklin Delano Roosevelt's meeting with King Abd al-Aziz Ibn Saud aboard an American ship in the Suez Canal in 1945 marks the beginning of a very important transition in the history of oil as a strategic commodity. Until the end of the World War II, the Middle East remained under British control. Nevertheless, as the war was drawing to a close, the United States decided to replace Great Britain as the leading Western military power in the region, fully aware not only of the wartime strategic importance of crude oil (recall that Hitler's two big defeats came in El-Alamein and Stalingrad, both on his way to the sources of oil, the Arabian peninsula and Baku, respectively) but also of the necessity of this strong source of energy for the postwar reconstruction (in Europe) and demand-led economic growth (in the States). In a sense, given the fact that the construction of new roads was a centerpiece of the New Deal even before the war (Yeomans, 2004, pp. 43–44), Roosevelt's courtship with King Ibn Saud and the oil-for-protection agreement between the United States and Saudi Arabia in the postwar era should not have come as a surprise. Today, the United States is the largest consumer of crude oil in the world, accounting for a quarter of the total consumption, followed by the European Union and China (British Petroleum, 2009).

The process of the gradual but secular growth of the U.S. demand for oil began in the 1920s, with the rapid adoption of motorcars as the individualized means of transportation and the first wave of suburbanization that it made possible. This "new mobile American way of life" came to an early grinding halt with the 1929 stock market crash and the onset of the Great Depression (Yeomans, 2004, p. 43). In this sense, the New Deal and the subsequent post–World War II Keynesian demand-led growth strategy transformed this incipient mode of organization of daily life into one of its central components. In the 1950s, the second wave of suburbanization and the construction

of the interstate highway system made the American auto industry the driving force of postwar economic growth.

This particular macroeconomic role of the suburban and mobile lifestyle made the demand for oil highly insensitive to changes in price. American consumers' dependence on oil was not limited to their demand for gasoline to fuel their cars with which they drive to work, to school, and to the shopping mall or to their demand for heating oil to keep their increasingly bigger suburban houses warm. The entire infrastructure of the mass production and transportation of (agricultural and industrial) consumer goods was also dependent upon increased mechanization and therefore upon increased demand for oil and oil-based products (e.g., petrochemical feedstocks, lubricants).

The relatively low price elasticity of demand for oil means that the consumers are highly dependent on oil in their consumption and cannot easily reduce their demand or switch to substitutes when its unit price increases (Cooper, 2003). Nevertheless, it is important to understand the historically dynamic nature of the price elasticity of demand. For instance, in the late 1970s and early 1980s, when the real price of oil increased dramatically due to supply disruptions caused by the Iranian Revolution in 1978 and the start of the Iran-Iraq War in 1980, the demand for oil declined by 16%. On the other hand, a similar increase in the price of oil during the 2000s did nothing to affect the steady growth of the U.S. consumption of oil (Hamilton, 2008). A possible explanation is that during the earlier price hike, consumers in the United States were able to substitute away from the nontransportation uses of oil, whereas in the current price hike, consumers had much more limited substitution possibilities in transportation uses of oil. In other words, the price inelasticity of the American demand for oil may be increasingly becoming a function of its suburban and mobile lifestyle.

Nevertheless, it is important not to reduce the growing and highly inelastic demand for oil to a mere effect of consumerist and materialistic "American way of life." Industrialization and the mechanization of the production process as a secular and global tendency throughout the twentieth century may be the underlying factor that drives the secular growth of the demand for oil. A number of commentators who investigate the matter from the perspective of Marxian political economy argue that the increasing reliance on nonrenewable sources of energy itself is an unintended consequence of the political and economic struggles of the working classes throughout the twentieth century (Caffentzis, 2008; Midnight Notes Collective, 1992). Viewed from this perspective, both the social Keynesianism of the post–World War II era and the increasing mechanization of the labor process throughout the century are seen as responses of the capitalist states and enterprises to the demands and resistance of the working classes (see also Mitchell, 2009). According to Marxian labor theory of value, human labor is the only new value-creating energy, and the capitalist system and its social institutions (the state, the corporations, the legal system, the ideological processes, etc.) have evolved through the past two centuries, in part, to manage and maintain the continual production, appropriation, and distribution of the surplus value by the working classes. These commentators, by noting the fact that "most energy derived from oil in capitalist society is involved in producing and transporting commodities and reproducing labor power [i.e., the consumption of consumer goods]" (Caffentzis, 2008, p. 318), argue that ever-growing demand for resilient and labor-saving nonrenewable resources in modern capitalist societies is an effect of the increasing difficulty of maintaining this system of exploitation.

Indeed, the gradually increasing demand for oil in newly industrialized countries such as China and India suggests that the driving cause may not simply be the "American way of life" but rather the class antagonism and the endless search for higher rates of surplus value extraction that propel both the increasing mechanization of the labor process and the increasing commodification of the reproduction of human capacity to labor. After the energy crisis of the 1970s, the U.S. economy went through a "neoliberal" restructuring, which led more and more American companies to outsource their production of consumer goods to developing countries such as China, where the value of labor power is significantly cheaper (Harvey, 2005). Today, a significant portion of the growing demand for oil in China (and other developing economies such as India and Brazil) is fueled by increased industrial production for the U.S. (and other advanced capitalist economies') consumer goods market (Gökay, 2006, p. 141). These cheap consumer products, in turn, made it possible for the increasingly precarious U.S. working class to continue to afford an increasing standard of living despite the fact that the real wages have remained stagnant since the early 1980s (Resnick & Wolff, 2006).

## Oil as a Strategic Commodity: Price Formation and "Scarcity" Maintenance

The standard neoclassical theory of the optimal pricing of exhaustible resources was established by Harold Hotelling in 1931 during a period when oligopolistic arrangements over the exploration, production, transportation, refinery, and distribution of oil were being struck both in the United States and in the Middle East to prevent the price of oil from falling below its cost of production (Bromley, 1991, pp. 95–98; Yergin, 1991, pp. 244–252). In this essay, Hotelling (1931) theoretically demonstrated that, under competitive conditions (and under very stringent, and unrealistic, assumptions pertaining to information, preferences, and technology), the price of an exhaustible resource, net of the cost of extracting the marginal unit of the resource, must grow along an efficient extraction path at a rate equal to the rate of interest (see also Devarajan & Fisher, 1981; Solow, 1974). While the price grows along the efficient

extraction path, the output (facing a stable demand) will decline asymptotically toward zero (Devarajan & Fisher, 1981, p. 66). Hotelling's analysis included a discussion of how under monopoly the price of oil will be initially higher but rise less rapidly and, accordingly, the depletion of oil reserves will be retarded. One important underlying presupposition of this analysis, as noted above, is that the price increases will "provoke changes in both production technologies and the consumption structure leading to substitution of the scarce resource with other, relatively less scarce, resource" (Roncaglia, 2003, p. 646).

While Hotelling's (1931) analysis is very useful in demonstrating the sheer impossibility of realizing competitive optimal outcomes in concrete, real economies, it has very little to offer in explaining the institutional complexity and historical trajectory of the political economy of the formation of the price of oil. In contrast to the abstract analyses of increasing "extraction" costs and optimizing firms found in the neoclassical tradition, the political economy approach offers an analysis of the oil industry as a multilayered process that involves the exploration, production, transportation, refinery, and distribution of oil (Bromley, 1991, pp. 87–90). In fact, it is possible to trace the history of the global oil industry and the changing institutional arrangements of price formation as different ways of organizing this multilayered process in response to shifting political, cultural, economic, and natural conditions.

Throughout its first century, the oil industry has been organized through the oligopolistic collusion of vertically integrated, large international oil companies (IOCs) that mobilized expert knowledge and expansive technology at all layers of this multilayered process, ranging from "upstream" activities such as exploration and production to "downstream" activities such as refinery, trade, and marketing. Under the "concession" system, which reigned until the early 1970s, IOCs used to purchase from sovereign states the rights to explore and exploit natural resources in return for fixed royalties. During the 1920s and 1930s, British Petroleum, various offshoots of the divided-up Standard Oil, and Royal Dutch/Shell controlled the oil industry, signing up "Red Line Agreements" among themselves to coordinate their activities in what used to be the Ottoman Middle East (Kurdistan, Iraq, Trans-Jordan, Arabian peninsula, and Gulf region; Yergin, 1991, pp. 184–206). But beginning with the 1950s, the anticolonialist, working-class struggles along with the emerging nationalist sentiments in oil-rich countries (e.g., antiracist labor strikes in Saudi Arabia, Baathist Arab socialism in Iraq, Iranian nationalism, Bolivarianismo in Venezuela) led to the formation of OPEC in 1960 in order to claim ownership of their resources and to enable oil-rich nation-states to exert greater control over the international oil market (Bromley, 1991; Mitchell, 2002, 2009; Vitalis, 2009). Nevertheless, the emergence of OPEC and the increasing role that national oil companies (NOCs) take in controlling the "upstream" of the industry did not necessarily lead to the demise of IOCs. On the contrary, as the industry shifted from an era of "free flow" to that of "limited flow" after 1974, they continued to remain highly profitable: They not only continued to account for nearly 10% of the net profits of the entire U.S. corporate sector (even better than their heyday in the 1930s), but also the rates of return of the large, U.S.-based IOCs remained above that of the *Fortune* 500 average—the dominant sector of the U.S. capital (Nitzan & Bichler, 2002, pp. 220–223). This is, in part, because IOCs have continued to work with the governments and the NOCs of the resource-rich nation-states by entering into upstream joint ventures and by continuing to control downstream business. But this is also because the costs of expanding the oligopolistic coalition are borne by the consumers of petroleum (as a means of both consumption and production): In comparison to the free-flow era (1920–1973), the average price of crude oil has tripled during the limited-flow era (1974–2008), from $15 to $45 (in 2008 dollars) (British Petroleum, 2009). Because resource-rich countries do not always have the necessary wherewithal to explore and exploit their resources, they tend to be dependent upon the expert knowledge and financial power of IOCs.

While oil production in an individual oil field, in contrast to the exploration for oil, tends to be predicated upon a basic level of technology, relatively high fixed costs, and economies of scale, neoclassical economists argue that the overall oil production at the level of the industry betrays increasing costs: "The more is produced, the more must one draw upon higher-cost sources" (Adelman, 1972, p. 5). Nevertheless, as noted above, historically, the areas that are exploited first have not necessarily been the easily accessible ones. In fact, the peculiar oligopolistic institutional configurations of the oil industry and the relatively extended periods of high prices have enabled relatively low-cost fields (such as those in Saudi Arabia due to its role as a "swing" producer) to remain underexploited while making relatively expensive fields (such as those in the North Sea, United Kingdom) economically viable for exploitation.

To appreciate why this is so, it is necessary to recall that, while there is a finite amount of oil under the surface of earth, within the myopic temporal horizon of the oil market, the problem historically has been the surplus, rather than the scarcity, of oil. The case of East Texas in 1930, when the price of a barrel of crude oil dropped to 10 cents per barrel as a result of uncontrolled competition and collapsing aggregate demand due to the Great Depression, is a well-known example of such cases of sudden flooding of the market with cheap oil in the absence of coordinated price fixing and sales regulation. Nevertheless, the surplus problem, or this tendency for overproduction in the oil industry, is more structural than it may initially appear. To begin with, at any given moment, given the nature of the industry, there is always an easily accessible excess reserve of oil

(both above and under ground). This gives the producers the opportunity and incentive to overproduce and increase their revenues in the short run by undercutting competition (e.g., governments that wish to finance an accelerated military buildup, the individual companies that wish to increase their market share or cash flow or both). In this precise sense, the historical trajectory of the price of oil can be read as a series of shifting institutional (oligopolistic) arrangements, based on a shifting and changing balance of power between petro-states and multinational corporations that collude to keep competitive impulses at bay.

The first thoroughly global configuration of the oil industry can be traced back to the 1930s. In the United States, the Texas Railroad Commission intervened in the East Texas "collapse" and divided the demand among the producers in proportion to their production capacity ("prorationing") and stabilized the domestic supply of oil. In the Middle East, the so-called Red Line Agreement of July 1928 and the infamous Achnacarry Agreement inaugurated the particular mode of oligopolistic arrangement that would regulate the allocation of the concessions among the Seven Sisters (Exxon, Mobil, Socal, Gulf, and Texaco from the United States and Royal Dutch/Shell and British Petroleum form Europe) until the 1970s. The oligopolistic domination of the Seven Sisters entailed the presence of a sizable surplus profit (monopoly rent) in the oil sector over a very long period of time. The concessions granted complete control over production across extensive areas for 60 to 90 years with complete control over pricing. And because the global price was determined according to the high-cost Texas crude as the benchmark, the Seven Sisters earned windfall profits from their low-cost production in the Middle East until the arrival of OPEC (Bromley, 1991).

Increasing demands for the nationalization of the local subsidiaries of IOCs that began in the 1950s and 1960s (e.g., Iran in 1951, Kuwait in 1960, Saudi Arabia in 1960, Iraq in 1964) culminated in the gradual nationalization of Aramco and a series of price increases in 1973–1974 in response to the 1973 Arab-Israel War and the United States' support of Israel. From this point onward, first Saudi Arabia and then OPEC as a whole began to play the role of the "swing" producer in the global oil market. A swing producer is defined by its capability to maintain an unused excess capacity of oil that can be switched on and off to discipline producers who may be tempted to undercut competition by producing above their allotted quotas. For an individual producer to be an effective swing producer, it has to have large enough and easily accessible (low-cost) excess capacity (Mitchell, 2002).

In the 1980s, in response to the price hikes of the late 1970s, the global demand for oil slumped, and OPEC, led by Saudi Arabia, played the role of the swing producer by cutting the production levels to adjust the global supply to the declining global demand. Nevertheless, as if making a demonstration of the structural tendency of the oil industry to overproduce, non-OPEC producers continued to increase their production rather steadily until stabilizing at 55% market share in 2004—even though OPEC continues to control three quarters of the proven oil reserves (British Petroleum, 2009). Today, Saudi Arabia continues to be the largest producer of oil, with 13.1% of the total oil production, and is followed by Russia (12.4%), a non-OPEC producer (British Petroleum, 2009). Some projections suggest that "total OPEC capacity is likely to fall significantly short—by the upwards of 5 million barrels per day—in the next decade" (Nissen & Knapp, 2005, p. 3) and that the Saudi Arabian production with 1.5 to 2 million spare capacity will not be able to make up for the difference, leaving "the global oil market with no institutional mechanism to control the upside of oil pricing" (Nissen & Knapp, 2005, p. 4). Nevertheless, significant evidence demonstrates that recent increases in the price of crude oil (U.S.$97 per barrel in 2008) cannot be simply explained by increasing global demand (its rate of growth has slowed down as the prices began to increase in 2005) or by supply problems or shortages (the proven oil reserves have been growing faster than consumption growth in recent years, and a number of low-cost substitutes, such as oil sands and oil shales, have become economically viable) (Hamilton, 2008; Wray, 2008).

A more realistic explanation suggests that the price increases are caused by "index speculation" that takes place in the futures markets for crude oil (Wray, 2008). Since the mid-1980s, actual negotiations and deliveries of oil contracts have been made based on the price of crude oil determined in spot and futures markets—where traders buy and sell futures contracts to either hedge against price fluctuations or to, plain and simple, speculate. The regional base price of the New York Mercantile Exchange (NYMEX) is represented by West Texas Intermediate—the type of crude that flows into the United States from its main ports on the Texas Gulf Coast. In contrast, the regional base price of the Singapore exchange is that of the Dubai crude. Even though the volume of trade in these markets may be very high, only a fraction of all trades ends up being realized, whereas most transactions are either "compensated" before expiration or rolled into newer futures contracts (Roncaglia, 2003, p. 655). While traditional speculative activity takes "the price risk that hedgers do not want," index speculators take only long positions by buying and holding a basket of commodities futures. Because these baskets are based on one of the commodity futures indexes (SP-GSCI and DJ-AIG), such speculative activities are named "index speculation" (Wray, 2008, pp. 63–64). But because these indices are based on the aggregation of different commodities (e.g., cotton, copper, corn, wheat, crude oil, natural gas) with varying weights (petroleum-related products account for 58% of the weighted average of SP-GSCI and DJ-AIG), the index speculators are insensitive to individual prices; they are only interested in the value of the index. As index speculation as an activity became popular (practiced by hedge funds, pension funds, university endowments, life insurance companies,

sovereign wealth funds, banks, and oil companies themselves), the volume of money that flowed into the indexes grew from U.S.$50 billion in 2002 to U.S.$300 billion in 2008, and along with the influx, the price of crude oil has increased dramatically (Wray, 2008, pp. 66–67). Without doubt, increasing prices may have also encouraged further index speculation—but it is important to acknowledge the role that speculative activity plays in determining the price of oil in the short run (Hamilton, 2008).

Increasing importance of spot and futures markets in determining the price of crude oil might give the impression that, provided that the speculative excesses of traders are regulated, the oil markets are becoming more and more competitive and the price is approximating toward the market-clearing equilibrium price. Yet, it is equally possible to interpret the increasing importance of futures markets both as a smokescreen to distract the general public from the enduring collusive arrangement between OPEC NOCs, IOCs, and the governments of oil-consuming, advanced capitalist economies, "allowing them all to bypass antitrust regulations," and as a mutually agreed upon mechanism for price formation, which would limit price competition among producers (Roncaglia, 2003, p. 656). To the extent that the supply of oil continues to be controlled by OPEC and the global demand for oil continues to grow at a secular pace, the futures markets, at their best, merely reflect these underlying oligopolistic forces (see also Hamilton, 2008).

Moreover, while the high oil prices driven by the speculative activity in commodities markets may bring windfall profits for both NOCs and IOCs, in the long run, high oil prices are not necessarily the best configuration for the economic interests of oil-producing economies either: Sustained high prices tend to provoke consumers to substitute away from oil-based sources of energy, slowing down the demand growth and rendering the market susceptible, once again, to overproduction. In fact, price instabilities caused by speculative activities have adverse effects on the macroeconomic stability of both net-exporter and net-importer economies. In short, the post-1970s reconfiguration of the oligopolistic control of the oil industry is not necessarily a stable one and requires continuing attention, maintenance, and management—if necessary, by means of military intervention and occupation (Moran & Russell, 2008).

Even the division of labor struck between NOCs and IOCs, where the former controls exploration and production and the latter refinery and distribution, is not a stable arrangement. Even though IOCs seemed to have survived the nationalization wave of the 1970s unscathed, retaining their profitability, they nonetheless produce only 35% of their total sales and own only a mere 4.2% of the total reserves. For this reason, they have continuing incentives, along with the U.S. government, which has historically supported them, to reestablish their control over the upstream end of the industry (Bromley, 2005, p. 252). In this regard, the new Iraqi Oil Law of 2007, which marginalizes the role of Iraqi National Oil Company by opening nearly two thirds of the oil reserves to the control of IOCs, constitutes an instance in which the IOCs and the U.S. government explore the possibility of tilting the balance of power within the post-1970s oligopolistic arrangement in their favor.

## Petro-States: Democracy, Economic Growth, and Class Conflicts

The term *petro-state* designates not only the "energy-surplus" oil-producing states but also "energy-deficit" oil-consuming states (Klare, 2008b; Mitchell, 2009). Economic growth and political stability of both producer and consumer states depend upon the uninterrupted and stable flow of oil between them. For the oil-producing states, the steady flow of oil provides a steady flow of revenues with which they can undertake public investments in infrastructures, purchase weapons and military technologies, induce economic growth through fiscal policy and transfer payments, redistribute income, invest in and incubate nonextractive sectors to replace the oil industry once the resources are depleted, or invest in international financial markets (sovereign wealth funds). For the energy-deficit advanced industrial states, the steady flow of "reasonably priced" oil can facilitate the smooth flow of transactions within the economy, sustaining a stable macroeconomic system, low unemployment levels, low levels of inflation, and sustained economic growth (in terms of the rate of growth of gross domestic product [GDP]).

Nevertheless, because there is a wide range of diversity among producer petro-states, it would be wrong to offer an ahistorical, general theory of state formation in petroleum-dependent economies. For instance, the dramatic failure of Nigeria's national project of petroleum-led development (Watts, 2006) cannot be lumped together with Venezuela's recent efforts to redistribute oil revenues to historically marginalized and impoverished sectors of the population (in particular, the urban poor and the indigenous populations). Similarly, while both the United States and China are energy-deficit petro-states that are dependent upon a steady flow of oil, the former is a global military power that has explicitly declared that it is ready to use "any means necessary, including military force" to protect its access to petroleum (the Carter Doctrine of 1980), whereas the latter is a fast-growing, export-oriented, state-controlled capitalist economy whose most important trade partner is the United States (Klare, 2004). Despite this internal diversity, it is still meaningful to suggest that a distinctive feature of the political economy approach to petro-states is to study the question of state formation in petroleum-dependent (producer or consumer) economies by analyzing the historical evolution of the political institutions, economic mechanisms, and social technologies as petro-states navigate, negotiate, govern, and manage their internal socioeconomic contradictions and class conflicts within the continuously realigning international geopolitical and economic context.

It is argued that energy-surplus countries tend to suffer from a deficit of democracy. Underlying this widespread perception is the assumption that oil revenues provide antidemocratic, authoritarian governments with the wherewithal to either buy off or repress political dissent (e.g., Ross, 2001). "Dutch disease" is the economic version of such "oil curse" arguments. Here the argument turns around the assumption that a booming natural extractive sector leads to stagnation or even deindustrialization in the manufacturing sector, leading to an imbalanced economic growth (e.g., Sachs & Warner, 1995). While such analyses of the "oil curse" may initially seem to capture some of the salient features of political economies of energy-rich oil states, they tend to obscure more than they reveal.

Political versions of the "oil curse" tend to represent the antidemocratic and authoritarian nature of these governments as a natural development, one that is bound to emerge given a presupposed natural human proclivity toward rent seeking in the absence of well-established property rights and competitive markets. Nevertheless, recent studies of the historical trajectories of state formation in producer petro-states suggest that antidemocratic, authoritarian governments emerge not because of oil revenues but rather to generate oil revenues in the first place. For instance, the emergence of Saudi Arabia as an authoritarian and sovereign oil state is intimately bound up in a history of repression of the antiracist, anticolonialist labor movement among petroleum workers (in particular, throughout the 1940s and 1950s; Vitalis, 2009) and the gradually increasing reliance of oil production on precarious immigrant workers (Midnight Notes Collective, 1992). Similarly, economic "Dutch disease" arguments tend to abstract from the international context within which economic policies are usually devised and implemented in many of the oil-rich yet underdeveloped economies. For instance, to be able to study the political economy of Nigeria as a failed state, it is necessary to look beyond the bureaucratic corruption and understand not only how the exploitation of oil from the Niger Delta has historically been based upon the oppression of indigenous peoples and cultures but also how the structural adjustment policies implemented throughout the 1980s and 1990s under the guidance of the International Monetary Fund (IMF) and World Bank have destroyed the social (multiethnic) and political (federal) fabric of the country by dismantling the welfare state through privatization and austerity programs (Midnight Notes Collective, 1992; Watts, 2006).

In contrast to energy-surplus countries, consumer petro-states appear to be much more democratic (with the exception of China). Nevertheless, there are two ways in which this assumption needs to be questioned. The first is the growing importance of economics and economic expertise in shaping the various aspects of the way states govern the "ordinary business of life." As discussed above, beginning with the Great Depression, the development and deployment of social Keynesianism (in the form of aggregate demand management policies) and New Deal liberalism (taking the shape of Great Society programs in the postwar era), which emerged as "a response to the threat of populist politics" during the 1930s, provided "a method of setting limits to democratic practices and maintaining them" (Mitchell, 2009, p. 416; see also Caffentzis, 2008–2009; Hardt & Negri, 1994). Second, throughout the postwar era, the United States consistently acted as an imperialist power conducting covert interventions in producer petro-states (hence shaping their formation) to protect the interests of the Seven Sisters (e.g., Iran in 1953, Iraq in 1963, Indonesia in 1965; Vitalis, 2009). As the Keynesian demand-led economic growth strategy (where wage increases followed productivity increases), along with the cold war geopolitical strategy of communist containment, became increasingly dependent on maintaining the free (and cheap) flow of oil through neocolonialist practices, the U.S. military began to gain increasing importance, gradually transforming the United States into an advanced national security state (Nitzan & Bichler, 2002). In the process, the sphere of democratic politics ended up being either usurped by the increasingly technical nature of economic expertise or regularly suspended by the concerns of national security (including ones pertaining to energy security).

As the era of free-flowing oil came to a close in the mid-1970s, the scope of democratic decision making in advanced capitalist social formations began to be limited with increasing vigor. The oil crisis came at a moment when the Keynesian regime of accumulation was not able to contain the working-class demands for an increased share of the social surplus (beyond the productivity increases), and the high price of oil quickly became an excuse for subsequent wage cuts (Caffentzis, 2008–2009). The attendant economic liberalism, which had been brewing at the Institute of Economic Affairs in London, the Mont Pelerin Society, and the Economics Department of the University of Chicago since the end of the World War II, emerged "as an alternative project to defeat the threat of populist democracy" (Mitchell, 2009, p. 417). Under the neoliberal regime of accumulation, the relationship between the state and the market was radically reconfigured, where the latter began to pursue a policy of active economization of the social life through marketization of social relations, privatization of the public sector, commodification of the commons, liberalization of trade, and financialization of daily life (Harvey, 2005). As life became more and more governed through market relations or market-based solutions, the postwar accord between the capitalist and the working classes broke down, wages ceased to increase in lock-step with productivity increases, and the tax cuts that were sanctioned by supply-side economics meant the dismantling of the welfare state and the reduction of government involvement in the economy to military Keynesianism (Resnick & Wolff, 2006). For the working classes of the consumer petro-states, the neoliberal deal meant, on one hand, stagnant wages, increasing work hours (and productivity), and increasing labor market

insecurity (the decline of full-time employment and the rise of precarious forms of labor) and, on the other hand, lower income taxes (but higher social security taxes), cheaper goods (trade liberalization), and increasing access to credit (financial deregulation) (Wolff, 2009). As if this was not enough to limit and diffuse the threat of democratic populism, after the attacks on September 11, 2001, and the subsequent invasion of Afghanistan and Iraq, neoliberalism took a neoconservative turn and further limited the sphere of democratic politics in the name of national security. To conclude, as we enter the twentieth-first century, given the fact that the petroleum-based modernization strategies of both producer and consumer states are in deep, structural crises, it may be useful to entertain the hypothesis that the "oil curse" is a disease that inflicts not only producer but also consumer petro-states.

## Conclusion: The "Real" Cost of Oil

Much of what has been discussed in this chapter so far has aimed at elaborating a political economy approach (as distinct from the standard neoclassical approach) to explain the concrete social and natural processes that make up the political economy of oil: the social construction of its natural limits; the social construction of the global oil demand; the social, economic, and political institutions that produce the price of oil; and the question of state formation in petroleum-dependent economies. An important assumption of the standard neoclassical approach is that, as the price of oil increases, over time, the world economy will gradually adjust its production technology and consumption patterns, substituting away from oil to alternative, less scarce resources. The political economy approach elaborated in this chapter suggests that there are a number of reasons why this may not be the case.

Let us leave aside for a moment the fact that the price of oil has historically been determined through oligopolistic arrangements (even when the buyers and sellers refer to spot and futures markets) and let us ask whether the windfall profits of the oil industry (shared between the oil-producing petro-states and their NOCs and IOCs) are invested in the research and development of viable alternatives to oil. Historically, petro-dollars have been extended as credits to developing countries (leading to the debt crises of 1980s), have enabled exploitation of more high-cost offshore fields (thereby delaying the need to develop alternatives), have been used to finance military buildups (the Middle East became the leading consumer of weapons and military equipment), have been used to invest in alternative business lines (e.g., the "financial sector" in Dubai, the "knowledge economy" in Qatar), and have been used to invest in financial markets (Davis, 2006). To say the least, none of these and other potential uses of the oil revenues necessarily facilitate the development of an alternative to oil. Moreover, the neoliberal tendency to try to solve all social and economic problems within the short-term horizon of market-based solutions makes it difficult to initiate and coordinate a concerted effort for the development of alternatives and the transformation of the production technology and consumption patterns at a global scale. Such a concerted effort requires a public recognition of the "real" costs of oil—namely, the human and real economic costs of energy wars, the ecological costs of the use of carbon-based sources energy, the social and economic costs of the distributional conflicts that are caused by climate change, the social costs of the "oil curse" both in producer and consumer petro-states, and the social and economic costs of economic crises that are triggered by the speculation-driven price of oil. For this precise reason, the main task of the political economy of oil in the twenty-first century should be to generate a widespread public recognition of the "real" costs of oil.

## References and Further Readings


Adelman, M. A. (1972). *The world petroleum market.* Baltimore: Johns Hopkins University Press.

Adelman, M. A. (1995). *The genie out of the bottle: World oil since 1970.* Cambridge: MIT Press.

Bowden, G. (1985). The social construction of validity in estimates of U.S. crude oil reserves. *Social Studies of Science, 15*(2), 207–240.

British Petroleum. (2009). *Statistical review of world energy.* Available at http://www.bp.com/statisticalreview

Bromley, S. (2005). The United States and the control of the world's oil. *Government and Opposition, 40,* 225–255.

Bromley, S. L. (1991). *American hegemony and world oil: The industry, the state system and the world economy.* University Park: Pennsylvania State University Press.

Caffentzis, G. (2008). The peak oil complex, commodity fetishism, and class struggle. *Rethinking Marxism, 20,* 313–320.

Caffentzis, G. (2008–2009). A discourse on prophetic method: Oil crises and political economy, past and future. *The Commoner, 13,* 53–71.

Cooper, J. C. B. (2003). Price elasticity of demand for crude oil: Estimates for 23 countries. *OPEC Review, 27*(1), 1–8.

Davis, M. (2006). Fear and money in Dubai. *New Left Review, 41,* 47–68.

Deffeyes, K. S. (2006). *Beyond oil: The view from Hubbert's peak.* New York: Hill and Wang.

Devarajan, S., & Fisher, A. C. (1981). Hotelling's "Economics of exhaustible resources": Fifty years later. *Journal of Economic Literature, 19,* 65–73.

Gökay, B. (Ed.). (2006). *The politics of oil: A survey.* London: Routledge.

Hamilton, J. H. (2008). *Understanding crude oil prices* (NBER Working Paper No. 14492). Cambridge, MA: National Bureau of Economic Research.

Hardt, M., & Negri, A. (1994). *Labor of Dionysus: A critique of the state-form.* Minneapolis: University of Minnesota Press.

Harvey, D. (2005). *The new imperialism.* New York: Oxford University Press.

Hotelling, H. (1931). The economics of exhaustible resources. *Journal of Political Economy, 39,* 137–175.

Hubbert, M. K. (1956, Spring). Nuclear energy and the fossil fuels. *American Petroleum Institute Drilling and Production Practice Proceedings,* pp. 5–75.

International Energy Agency. (2004). *World energy outlook.* Available at http://www.worldenergyoutlook.org/2004.asp

Klare, M. (2004). *Blood and oil: The dangers and consequences of America's growing dependency on imported petroleum.* New York: Metropolitan Books.

Klare, M. (2008a). Petroleum anxiety and the militarization of energy security. In D. Moran & J. A. Russell (Eds.), *Energy security and global politics: The militarization of resource management* (pp. 38–61). London: Routledge.

Klare, M. (2008b). *Rising powers, shrinking planet: The new geopolitics of energy.* New York: Metropolitan Books.

Midnight Notes Collective. (Ed.). (1992). *Midnight oil: Work, energy, war: 1973–1992.* New York: Autonomedia.

Mitchell, T. (2002). McJihad. *Social Text, 20*(4), 1–18.

Mitchell, T. (2009). Carbon democracy. *Economy and Society, 38,* 399–432.

Moran, D., & Russell, J. A. (2008). Introduction: The militarization of energy security. In D. Moran & J. A. Russell (Eds.), *Energy security and global politics: The militarization of resource management* (pp. 1–17). London: Routledge.

Nissen, D., & Knapp, D. (2005). Oil market reliability: A commercial proposal. *Geopolitics of Energy, 27*(7), 2–6.

Nitzan, J., & Bichler, S. (2002). *The global political economy of Israel.* London: Pluto Press.

Resnick, S. A., & Wolff, R. D. (2006). *New directions in Marxian theory.* London: Routledge.

Roncaglia, A. (2003). Energy and market power: An alternative approach to the economics of oil. *Journal of Post Keynesian Economics, 25,* 641–659.

Ross, M. L. (2001). Does oil hinder democracy? *World Politics, 53,* 325–361.

Sachs, J. D., & Warner, A. M. (1995). *Natural resource abundance and economic growth* (Development Discussion Paper No. 517a). Cambridge, MA: Harvard Institute for International Development.

Solow, R. M. (1974). The economics of resources or the resources of economics. *American Economic Review, 64*(2), 1–14.

Vitalis, R. (2009). *America's kingdom: Mythmaking on the Saudi oil frontier.* London: Verso.

Watts, M. (2006). Empire of oil: Capitalist dispossession and the scramble for Africa. *Monthly Review, 58*(4), 1–17.

Wolff, R. (2009). *Capitalism hits the fan: The global economic meltdown and what to do about it.* New York: Olive Branch Press.

Wray, L. R. (2008). Money manager capitalism and the commodities market bubble. *Challenge, 51*(6), 52–80.

Yeomans, M. (2004). *Oil: Anatomy of an industry.* New York: New Press.

Yergin, D. (1991). *The prize: The epic quest for oil, money, and power.* New York: Simon & Schuster.

# 64

# Transportation Economics

Anthony M. Rufolo
*Portland State University*

To begin, transportation economics covers a broad range of issues. There are differences between personal and freight transportation, among transportation modes, and between the fixed and variable parts of the transportation system. Analysis of personal transportation tends to focus on commuting and choice of mode, although there is also substantial interest in other personal transportation choices, such as time of day for travel and grouping of trips. For freight, the issues are primarily related to the cost of freight movement and the damage that heavy vehicles do to roads. Each mode has similarities and differences in the issues to be analyzed. For virtually all transportation systems, there is a large, typically publicly owned, infrastructure and a variable, often privately owned, set of vehicles that use that infrastructure (Dean, 2003). The financing systems are often complex, and the incentives that they offer may generate further issues that need to be addressed in evaluating transportation systems.

In discussions of transportation, there is near universal agreement that there are problems. Each mode has different problems. For automobiles, increasing congestion, difficulty in financing construction and maintenance, and concern over the environmental effects are major issues. For public transportation, a long-term downward trend in share and rising costs are the key concerns. For freight, competition and cooperation among the modes, capacity of the freight system, and allocation of cost are among the important topics.

This chapter starts with a discussion of the demand for transportation services and the factors that are important in analyzing the choices made from the perspective of the user of the transportation system. Then, of course, supply of transportation is evaluated. Because the market for transportation services is not the typical market system, the method of funding and the incentives for efficient use of the system are then discussed, along with the role of government in regulating the transportations system. The chapter concludes with a discussion of some of the important policy debates regarding transportation.

## Demand for Transportation Services

The trend for personal travel has been for increasing use of the automobile and reduced reliance on alternative modes. One result has been increased levels of congestion and delay on the road system and increasing subsidies for the transit system. From an economic perspective, many of the perceived problems occur because people do not pay the appropriate price for travel. Hence, it is important to understand the demand for transportation and the methods of finance because the latter determines the perceived price.

There are two important distinctions between the demand for transportation and the demand for most goods and services. The first is that transportation is typically classified as a derived demand; most travel is not consumed for itself. Rather, it is a method to achieve other goals. The second is that for personal transportation, the person's time must be used, and the value of this time is part of the cost of transportation. Thus, time and the value of time are very important issues in discussions of transportation. In fact, transportation economists often differentiate between the cost of transportation services and the cost of transportation, which includes the opportunity cost of the time used in making the trip. The latter is typically referred to as *generalized cost.* This distinction is very important when analyzing the demand for transportation

services because the differences in time cost often have substantial impacts on the choice of mode for travel.

The next step in analyzing the demand for personal transportation refers to elasticity of demand. In addition to the common discussion of price elasticity of demand, the income elasticity of demand and cross-price elasticity of demand are important in analyzing transportation choices.

## Cost and the Value of Time

The most common example of the distinction between the generalized cost and the monetary cost of different choices is the choice of mode for commuting. If one looks only at the monetary cost, then mass transit would be a bargain, compared to driving, for most people. The transit fare is typically a fraction of the cost of using an automobile, especially if the person driving must also pay for parking. Despite the price differences, the vast majority of commuters in the United States choose the automobile over mass transit. A major reason is that the auto commute is typically much shorter than the transit commute, and people value the time savings. A broad generalization often used in transportation analysis is that people value time in commuting at about half of their wage rate, but there is substantial variation in people's willingness to pay to save time. This valuation also varies by how the time is being used. Time spent waiting for a transit vehicle costs more than time spent in the vehicle, and time spent driving in congestion is viewed as being much more costly than time driving at free flow. More recent research also finds that there is considerable variation across individuals in the value they place on time (Small, Winston, & Yan, 2005).

Another aspect of the value of time is the reliability of the system. Although average travel time is very important, the variation in travel time may be equally important to many people. Getting to the destination either early or late may impose costs on the traveler. The cost might be being late for work or a meeting, or it might be having extra time before work starts. The greater the variation in travel time, the more of a cushion is needed to be reasonably certain of getting to the destination at a specific time. One method of describing this is to look at the probability distribution of trip times. For example, the average trip time may be 20 minutes, but because of wide variation in congestion, the traveler may have to leave 30 minutes before the desired arrival time to have a 90% probability of arriving on time. In general, higher levels of congestion are associated with higher levels of uncertainty in addition to the longer travel time. This may make it difficult to identify the value of time for the traveler, and there is evidence that people place a separate value on improved reliability compared to reduced travel time (Lam & Small, 2001).

Time is also of importance in freight transportation, although it is typically less a factor than for personal transportation choices. Some items are perishable, so the value of time is obvious, but the use of overnight express and similar services show that saving time in the movement of freight can also be valuable. In addition, many firms have come to rely on timely delivery of inputs as a method to reduce the need to hold large inventories of the items that they use in the production process. Because late delivery can disrupt the production process, both time and reliability are important for these freight services.

## Elasticities

There are a variety of elasticities that are important in understanding transportation economics. The price elasticity of demand is the one most commonly discussed in economics, and it is very relevant for transportation. However, two other elasticities are important in analyzing transportation choices: the income elasticity of demand, which relates to changes in demand as income changes, and the cross-price elasticity of demand, which relates to the way demand for one good changes when the price of another good changes.

### *Price Elasticity*

For transportation, the price elasticity of demand is complicated because there is either the ordinary price elasticity of demand, based on monetary price, or the generalized cost elasticity of demand, based on monetary and time cost. Where the time needed for travel does not change, the ordinary price elasticity of demand is evaluated. In general, the price elasticity of demand for transportation is fairly low in the short run; people do not appear to be overly sensitive to price in deciding whether to make a trip. For example, the general rule of thumb is that the price elasticity of demand for transit is about 0.3. Hence, a 10% increase in transit fare is expected to lead to a 3% reduction in the number of riders. However, the estimates of the elasticity of demand cover a wide range (Holmgren, 2007). Elasticity of demand for a particular mode or at a particular time may be higher or lower than the average because there will be different opportunities to substitute other modes, routes, or times, and the availability of substitutes makes demand more elastic. A variety of other price elasticities are sometimes discussed, such as the elasticity of demand for gasoline. The elasticity for gasoline is found to be very low in the short run (Congressional Budget Office, 2008), but this is largely because the cost of gasoline is only a part of the cost of the trip. Any given percentage increase in the price of gasoline will be a much smaller percentage of the cost of the trip, and the full cost of the trip is the more relevant consideration.

### *Income Elasticity*

Another very important elasticity is the income elasticity of demand. It compares the change in the demand for a good or service at fixed prices with the change in income.

The demand for most goods and services is expected to increase with income. However, if the demand decreases as income increases, it is labeled an inferior good. Generally, inferior goods are goods that have a higher quality substitute available, and as income rises, people shift to the higher quality substitute.

In transportation, one sees that the demand for automobiles is highly income elastic. Both within countries and across countries, one sees rising demand for automobile ownership as income rises. This, of course, affects the demand for using transit, which is typically seen as an inferior good. There are many factors other than income that affect the demand for different types of transportation, but the general trend is consistent. The demand for more transit increases over some income ranges, but over most levels of income, the demand for transit decreases as income increases, everything else constant.

*Cross-Price Elasticity of Demand*

Another important elasticity concept is the cross-price elasticity of demand. It refers to how the demand for one good changes when the price of another good changes. This is the typical method to identify complements and substitutes. If the price of one good rises, people tend to use less of it. If as a consequence they buy more of some other good, that other good is a substitute. However, if the goods are complementary, then the reduction in the purchase of one good would also lead to a reduction in the purchase of its complements. Transit advocates argue that providing more transit services or reducing transit fares would reduce the number of automobile trips. This is an argument that the two are substitutes, and the amount of substitution then depends on the cross-price elasticity of demand. This is an empirical parameter, and it will differ among areas and over time. The evidence in the United States is that in most cities, the cross-price elasticity of demand between transit and automobiles is very low, if not zero. This means that lowering transit fare does very little to get people out of their automobiles. As noted earlier, the price elasticity of demand for transit is fairly low. This implies that lowering fares is not extremely effective in getting people to use transit, and some of the increase in transit usage is caused by an increase in trips taken or diversion from other modes, for example, carpooling, walking, or biking, rather than a diversion from trips in automobiles.

The other factor of importance with respect to cross-price elasticity is the relative shares of the substitutes. For example, if transit carries 10% of the trips and automobiles carry the other 90%, then even if all new transit trips represent a shift from automobiles, a 10% increase in transit usage (to 11%) would decrease auto use by only about 1% (to 89%). This means that transit-oriented policies are likely to have noticeable effects on auto use only in areas that already have large amounts of transit use.

## Supply of Transportation Services

Just as the analysis of the demand for transportation had to take account of the unique characteristics of transportation services, the supply analysis is affected by them as well. Much of transportation infrastructure is large scale. In the United States, roads, transit, and airports are typically provided by the public sector, although there is increasing interest in private ownership and provision. Economies of scale in providing the infrastructure and the combination of passenger and freight transportation are two of the more important issues in the supply of transportation services.

### Economies of Scale

Economies of scale relate to the relationship between the cost per unit and the number of units being produced. Economies of scale are often important in transportation, but the focus is typically somewhat different than for most goods and services. For example, mass transit requires many users for it to be cost effective. However, there are different ways to measure scale; one is simply system size, and the other relates to service over given segments of the system. The distinction is between the size and the density of the network. Density relates to the number of people wanting to make a particular trip, typically from a common set of origins to a common set of destinations. As more people want to make the same trip, the cost of providing that trip per person is often reduced, although the range of scale economies may be somewhat limited. If one considers simply the size of the system, then adding more routes would increase scale, but it is less likely there will be economies of scale using this measure.

In addition to the number of people wanting to make essentially the same trip, one also must consider the options that are available for a trip. In this case, the network becomes an important consideration. The network refers to the places that one can get to on the transportation system. For the automobile, the network is essentially any place that has a road, but for other transportation systems, the network is typically much more constrained. For a mass transit system, one might consider the network to be simply the areas within easy walking distance of stations or stops.

Another way to think about network effects is to consider each trip as composed of three separate functions. These are typically defined as collection, line haul, and distribution. First, the passenger must get to the transportation system. Then there is typically a relatively high-speed movement to some other point in the system, from which the traveler must get to the ultimate destination. For air travel, the three functions would be getting to the airport, the plane trip, and then getting to the ultimate destination. For commuting by car, each segment tends to be done in the vehicle, although there may be a walk to the vehicle at the beginning and the end. For

transit, the collection phase may involve simply walking to a transit stop, but it may also involve driving or biking to a transit stop or using a feeder service to get to the main transit stop. Historically, collection was typically associated with walking to transit, and many residential areas developed around streetcar or other transit services. Once the passengers are on the transit vehicle, the majority of the distance is covered. This may be on local service or some form of express service. Finally, when the passengers leave the transit vehicle, they must get to their destination, typically by walking.

### Economies of Scope

Another issue in the cost of providing transportation services is the ability to provide different types of transportation services. Using the same facilities to provide different types of transportation service is called *economies of scope.* The largest distinction in this area is between personal transportation and freight transportation. For example, both automobiles and heavy trucks typically use the same roads. This makes sense only if there are economies of scope in the provision of roads, and most studies conclude that this is indeed the case. In other words, it would be possible to have a separate network of roads for trucks and for automobiles, and occasionally there are such separate facilities. However, they are rare. There are disadvantages of mixing automobile and truck traffic on the same road. For example, roads must be built to higher standards to withstand the damage done by heavy vehicles, so a road built solely for automobiles could have thinner pavement. Safety concerns, speed differences, and the discomfort some drivers feel near large vehicles are also associated with mixing the vehicles. On the other hand, there are benefits to mixing the traffic. Roads with two lanes in each direction can carry more than twice as many vehicles as roads with only one lane in each direction because they allow easier passing and other operational improvements. Separate roads would also have to have separate fixed costs, like shoulders. It also seems that there are some benefits related to time of usage, with automobiles having more usage during peak congestion periods and trucks often showing greater flexibility to use the roads at less congested times (de Palma, Kilani, & Lindsey, 2008).

Passenger rail and freight rail are much less compatible. For passenger rail service, time is very important, but for freight rail service, the emphasis is on keeping the cost low. Hence, freight trains are often large and slow moving. Because it is difficult to pass another train, passenger and freight traffic tend to interfere with each other. If the volume of traffic is low, the economies of scale in sharing track may make combined service the least costly option, but as volume increases, the diseconomies of scope typically cause separation of the activities.

## Mode Choice

Mode choice is important both for personal travel and for freight movement. Mode choice for commuting has received substantial interest because of the growth in the share of single-occupant vehicles and the decline in share for transit, carpooling, and other modes. Although the mode shares differ substantially across countries, the trend tends to be fairly universal. To some extent, this is the result of rising incomes and people placing a higher value on saving time, but there are also substantial concerns about whether people are making those choices based on full information regarding the cost differences.

### Location Patterns

If one thinks about the commute in terms of the collection, line haul, and distribution phases, it becomes apparent that changes in location patterns over time have had a substantial impact on the ability of transit to serve these functions for commuters. As population decentralized, development moved away from concentration around transit stops, but transit could still serve effectively if people could get to the system and employment was concentrated around transit. Hence, park and ride became a viable option for people with a car available. The increased ownership of automobiles over time made this a possible method to use transit for many people. However, the decentralization of employment has proven to be more problematic. Although people are willing to take cars that were purchased primarily for personal use and let them sit in a parking lot while at work, they typically are not willing to purchase a car to be used to get from the transit station to work. Hence, a commuter often can still use transit if his or her residence is not located near a transit stop, but it is more difficult if his or her employment is not near a stop. Hence, employment location patterns have become an important issue for the viability of using transit.

### Components of Cost

Another important consideration in mode choice is the cost perception on the part of the person making the decision. Economic theory tells us that the efficient decision depends on the decision maker's facing of the full marginal cost. However, it is seldom the case that the person making a commute decision will face that cost. The distinction between fixed cost, marginal cost, and external cost is helpful in understanding the potential distortion. In making a decision between auto and transit commuting, the person is likely to take account of both fixed and variable costs, but once the choice to use an auto is made, only variable costs enter the decision for a particular trip. Further, certain costs may be paid indirectly or by someone else and will not enter into the decision.

If one does not own a car for other reasons, then the full cost of purchase and maintenance will enter into the mode choice decision. Once the automobile is purchased and other fixed costs, such as insurance, are paid, only the variable cost of using it for a particular trip will be taken into account. The income elasticity of demand for automobiles has affected the relative cost of using transit and cars based on the marginal cost comparison. As income increases, people who are transit users may still find that the automobile is very convenient for recreation and household use. Yet once the vehicle is purchased, the cost of using it for commuting is only the variable cost, and this will shift the commute mode choice decision. There have been some experiments associated with converting some fixed cost of auto use into variable mileage costs to see how behavior would change (Abou-Zeid, Ben-Akiva, Tierney, Buckeye, & Buxbaum, 2008), and people do seem to drive less when the mileage charge is higher (Rufolo & Kimpel, 2008).

Aside from the difference between fixed and variable costs, automobile users, especially at peak times, do not pay all of the cost associated with automobile usage. Some costs are simply paid indirectly or by others. Parking is often cited in this category because relatively few employees pay for parking at their places of work (Shoup, 2005). Other costs are paid by the drivers, but those costs do not reflect the full marginal cost. Congestion falls into this category because the cost of increased delay with congestion is imposed on drivers, but the marginal cost of one more driver exceeds the average cost paid by the driver. This complex subject is covered in detail in a later section. Finally, automobile usage generates substantial negative externalities in the form of pollution and related costs, and these costs are not paid by automobile users (Parry, Walls, & Harrington, 2007).

Transit users typically do not face efficient prices either. Large subsidies keep the fare charged substantially below the marginal cost of providing service, although a later section shows that there is some disagreement about the optimal transit subsidy. Peak-period transit fares typically should be higher than off-peak fares to reflect the fact that the need for capital stock is determined by the peak usage. In addition, fares should be higher for longer trips, and there should be a variety of other adjustments to reflect cost differences. Few transit agencies follow any of these principles, so commuters by transit also tend to pay substantially less then the optimal charge. One effect of the pricing system is that there is likely to be too much consumption of all transportation services relative to the optimum.

## Congestion

Road congestion is a significant and growing problem in most countries. Economists typically define the effect of congestion as an external effect of using the transportation system. There are actually several different causes of congestion. The first relates to bottlenecks. Bottlenecks reflect a reduction of capacity. A reduction in the number of lanes can certainly create a bottleneck; however, in transportation, they can occur for a variety of reasons. For example, places where traffic enters and leaves a limited-access road may have reduced capacity even though the number of lanes is unchanged, and the bottleneck is defined by the reduction in capacity. Next, there is congestion caused by incidents. These may be accidents or simply stalled vehicles or animals on the road. Finally, there is systemic congestion. This occurs based on the number of vehicles trying to use a road, and it can best be thought of as an effect that each driver has on all other drivers. For safety, there must be some distance between vehicles. As more vehicles try to use the same road, the distance between vehicles becomes compressed. The crowding forces traffic to slow. Because the slowing is associated with the number of vehicles on the road rather than some vehicles slowing others, it is typically analyzed as being caused equally by all users of the road.

Economists have concluded that there is a substantial difference between the additional time that each driver must take and the effect that having one more vehicle on the road creates for the whole system. To help see this, consider the following example. Suppose that if 1,000 cars per hour try to use the road, there is completely free flow, and each driver takes 10 minutes to complete his or her trip. However, if 1,001 cars per hour try to use the road, the slight slowing causes each driver to take 10 minutes and 1 second for the trip. The total travel time for 1,000 cars is 10,000 minutes, but the total travel time for 1,001 cars is slightly more than 10,026 minutes. Thus, although each driver sees a travel time of a little over 10 minutes, the total travel time for all drivers has increased by over 26 minutes. Because any one of the drivers could reduce the total by 26 minutes by not making the trip, it is clear that none of them are considering the full effect that their using the road has on the entire system.

### Congestion Pricing

The effect of most types of congestion is that the cost to individual drivers is less than the cost to the system. As the example shows, someone who values the trip enough to make it if the time cost were as high as 15 minutes would, if faced with the full 26-minute cost imposed on the system, decide not to make the trip. Economists propose that all drivers be charged the difference between the cost that they face and the cost that their use of the system imposes on the whole system. This is known as congestion pricing.

It can be formally demonstrated that the idealized system of pricing would generate net benefits for society through more efficient use of the road system. However, drivers have been strongly opposed to such pricing systems. The basic reason is that to make the system work,

most drivers must be made worse off than they are with the so-called free system. It may seem paradoxical that there is an improvement if most drivers are worse off. The reason is that under congestion, drivers pay with time, but under a toll, they pay with money. The time that they waste in congestion is a cost to them but provides no benefit to anyone. The toll that they pay is also a cost to them, but it is just a transfer of money to the toll agency. This money can be used to lower other taxes or provide additional benefits. Hence, the individual drivers are worse off, as they are with any tax or fee, but the transfer of money allows for some offsetting benefits to be created, while the wasting of time does not.

There are a variety of types of congestion pricing. State Route 91 (SR91) in California has two lanes in each direction that are priced and four lanes in each direction that are unpriced. The price is changed as often as every hour, but the rates are set in advance. The disadvantage of prices set in advance is that the demand for using the road varies randomly to some extent. The prices that are set in advance may have to be high on average so as to generate free flow most of the time. This could lead to less usage than would be efficient. With the fixed charges, drivers then have the choice of paying and saving some time or not paying and facing congestion. Studies of the usage of SR91 find that relatively few drivers use it every day, with many drivers using it occasionally. The explanation is that it may be worth it to save time on some days but not on others, for example, if parents are late to pick up their children.

Interstate 15 in California has two reversible lanes that are free for high-occupancy vehicles (HOV) and charge a price that is varied as frequently as every 6 minutes for other cars so as to maintain free flow. The benefit of the dynamic pricing is that it allows more vehicles to use the lanes in periods of low demand while still maintaining free flow during periods of very high demand. The disadvantage for drivers is that they do not know until they arrive at the entrance what the price will be. The price is displayed on an electronic sign, and they have a short time to decide whether to take the priced road or to stay on the free one.

A number of cities have adopted a system of charging vehicles either for entry into an area or for any driving in that area. Singapore is widely cited as the first city to use this system, but London has received substantial interest for instituting its system. London charges a flat fee for any vehicle that drives within the designated zone and designated times. The fee varies with type of vehicle and for a variety of other reasons, but the basic fee is a flat charge. The fee is enforced with a system of video license plate recognition.

There is one additional drawback to congestion pricing. It costs money to collect and administer the charge. For example, one study of the London system concluded that the cost to administer the system was so high that it offset the benefits from better traffic flow (Prud'homme & Bocarejo, 2005). Reductions in the cost of the equipment needed to impose charges, along with improvements in the administrative capability to collect revenue, cause these issues to decline in importance over time, but they must be taken into account when evaluating the net benefits of using pricing to manage congestion.

## Hypercongestion

Congestion can become so severe that there is actually a reduction in the number of cars that get through to their destinations in each time period. The easiest way to illustrate that would be to think of complete gridlock, the ultimate congestion. In this case, no vehicle gets to its destination. There is some disagreement about when congestion gets bad enough to cause an actual reduction in the flow of vehicles. For many years, it was thought that a speed of about 30 to 35 miles per hour maximized flow, but recent research suggests that flow may be reduced when speeds drop below free flow. Where hypercongestion exists, congestion management has the potential to increase both speed of travel and the number of vehicles traveling (Varaiya, 2005).

## Congestion Policy

Although economists recommend pricing that varies by time of day to manage congestion, there is relatively little support for this approach. Most congestion policy relates to other methods to relieve congestion. Congestion caused by too many people trying to use the road can be addressed only by changing their demand to use the road. Pricing is the most effective way to do this, but other types of demand management also can be effective. The most common is the use of ramp metering to manage the number of vehicles entering a restricted access road. Ramp meters can improve the flow on the metered roads, but they have some drawbacks as well. They can cause backups onto surface streets, and they favor vehicles making long trips over vehicles making short trips. Responses to other types of congestion may differ. For example, the best response to congestion caused by incidents seems to be to work to rapidly clear the incident. Many states now have operations to do precisely this. They encourage people in fender benders to move to the side of the road and may have service vehicles to help clear accidents or stalls.

## Latent Demand

As the time cost of a trip increases, people make various adjustments to their travel plans. With respect to peak period congestion, the term *triple convergence* is often discussed (Downs, 2004). Anthony Downs argues that when faced with increased congestion, some people respond by changing their time of travel to less congested periods, others change mode of travel, and still others change the route of travel. Another possibility is to choose not to make the trip. These changes reduce the maximum peak congestion from what it would be if people had not changed their

behavior, but it also means that adding capacity has less impact on the maximum amount of congestion than it would if people did not change their behavior. As congestion decreases, people will shift back to their preferred times, routes, and modes. Hence, the triple convergence then offsets some of the benefits of the increased capacity.

With triple convergence, people may now be traveling closer to their ideal times, taking a more direct route, or traveling by a preferred mode, but these shifts then mitigate the effect on peak congestion that the increased capacity generates. To be sure, there are substantial benefits to these shifts, and the period of peak congestion is likely to be smaller. However, some see these shifts and argue that there is no benefit to building additional road capacity because it simply gets used up. They term the increase in usage as being *latent demand.* Although the convergence of travel times does mitigate the benefits of the capacity expansion, it is a serious mistake to conclude that there are no benefits. Nevertheless, understanding triple convergence and latent demand gives us more capability to accurately predict the impact of transportation investments.

## Mass Transit

The long-term trend for transit is a declining share of personal transportation. Most transit in the United States was privately owned and largely funded from fares until about the 1960s. Since then, most systems have been converted to public ownership and rely heavily on public subsidies for funding. Critics contend that this has caused substantial inefficiencies in transit operations (Lave, 1994), while supporters argue that transit provides offsetting benefits. Hence, the important economic concerns are how the organization of transit affects efficiency and the arguments for public subsidies.

### Economics of Transit

As noted earlier, two key issues for transit are the economies of density and the network characteristics. John Meyer, John Kain, and Martin Wohl (1965) are credited with first analyzing the relative cost of different methods of urban transportation. They did not consider the value of time, so they were just considering monetary cost. Their conclusion was that the automobile was the low-cost alternative for low density of use and that buses were then the low-cost alternative, except for very high density, when rail would have the lowest cost. Although actual density has increased in most cities over time, the density of transportation demand often has not. The transit system economies occur when a large number of people want to make the same trip at the same time. Decentralization of first residences and then employment has often reduced the density of trips, making transit more costly despite the increase in overall density. In addition, the outward spread of most urban areas has reduced the percentage of residences and places of employment that are within the transit network. Finally, the increased value of time associated with rising incomes has increased the generalized cost of transit relative to auto travel.

The issue of public versus private provision generates substantial controversy. A number of other countries, including England, have moved toward more privatization of their transit systems. In the United States, some transit systems contract with private providers for service. Supporters of privatization argue that the competition leads to lower cost and improved efficiency, and the evidence seems to support this conclusion.

#### *Transit Fares*

Transit has peak demand that is similar to that for automobiles. Economists typically argue that prices should be set to reflect the cost of providing the service. From an economic perspective, the cost of providing peak transit service is higher than the cost of providing off-peak service. This seems counterintuitive to most people. The large number of people using the system during the peak means that the cost of running the vehicle is spread over more passengers than in the off-peak hours, but the number of vehicles needed by the system is determined by the peak usage. Thus, more of the cost of the system is attributed to the peak than to the off-peak usage.

The counterargument is that each person who uses transit actually creates a benefit for other users (Mohring, 1972). The basis for this conclusion is that as more people use the system, more service is provided and average wait time decreases. Average wait time for transit is somewhat dependent on the frequency of service. If people arrive randomly at the transit stop, then the average wait is half of the time between vehicles. So increasing the number of vehicles leads to reduced average wait time, and this means that the efficient fare would be lower than the marginal cost of providing the service because as more people use the system, wait time for others is reduced, creating an external benefit that justifies a subsidy. Despite this argument, Charles Lave (1994) finds that the subsidies have largely resulted in inefficient production rather than more service.

Although economists argue about whether subsidies promote or hinder efficiency in transit, the popular argument for transit subsidies is that the lower fare can be used to entice people out of their automobiles. However, the evidence on very low cross-price elasticity of demand between transit and automobiles means that this argument is largely incorrect for most cities.

## Regulation

There is a long history of regulation of the transportation industry. Many types of regulation relate to issues such as safety, but there has also been substantial economic

regulation of the various modes. For many years, the federal government set prices for airlines, railroads, and trucks. Although the intent was to protect consumers, the effect of deregulation of these industries has been substantial improvements in productivity and reduction in prices (Winston, 1998). To be sure, many of the improvements are viewed negatively by some. For example, under regulation, railroads were required to maintain substantial amounts of service where the cost exceeded the revenue, but they were then able to compensate by charging higher prices on service in high-demand areas. Under deregulation, prices declined in the high-demand areas, and much of the service to low-demand areas was discontinued. This is an improvement in efficiency, but it is negative from the perspective of those losing service.

Taxi regulation is one area of transportation regulation where there does not seem to be much prospect for reform. Many cities restrict the number of taxi licenses that they grant. From an economic perspective, the restrictions on entry are likely to cause prices to increase and service to be concentrated in the most profitable areas. More competition is expected to improve service and result in reduced prices. Yet experience with taxi deregulation has been problematic. This may be because certain types of competition are not allowed in the deregulated markets. For example, most airports require that passengers take the taxi at the front of the line. There is no opportunity for one farther back to offer the service at a lower price. Hence, deregulation may still not allow much competition. One study concludes that a whole new regulatory structure is needed for all parts of the transit and taxi system (Klein, Moore & Reja, 1997).

## Issues

Funding for transportation has received substantial attention. For example, Congress created two separate commissions to study and make recommendations on transportation finance: the National Surface Transportation Policy and Revenue Study Commission and the National Surface Transportation Infrastructure Financing Commission. Each concluded that existing finance mechanisms were insufficient and recommended changes. The method of funding and the level of funding are both sources of controversy. In personal transportation, there are disputes about whether auto users or transit riders pay the appropriate costs. In addition, there is substantial concern that both highways and transit face substantial challenges with respect to finance.

### Highway Finance

In an economist's ideal world, prices for vehicles using the road system would be set to reflect the cost the vehicle imposes on the system (Winston & Shirley, 1998). The economic system works most efficiently when price is equal to marginal cost. For cars and other light vehicles, the primary determinant of the efficient price would be the level of congestion on the road in use. Where congestion is heavy, the price would be high, and a low price would be charged for travel on uncongested roads or during off-peak times. The high price would discourage use during the peak (Rufolo & Kimpel, 2008) and induce more use of alternative modes, including carpooling. The substantial decline in the number of true carpools has caused some people to argue that carpooling will not occur because people value their time highly and carpools impose a time cost for formation. However, there has been spontaneous carpool formation where there is an incentive, such as reduced travel time for those in carpools (Spielberg & Shapiro, 2001). The revenue generated from pricing would also serve to guide new investment. Where revenue is high, the value of added capacity will also be high, so the price serves as a sign that more investment should be considered.

For heavy vehicles, the price should vary with the road damage done and congestion. Heavy vehicles do substantially more damage to roads than light vehicles, and the damage is largely related to the weight per axle of the vehicle. Oregon charges a weight-mile tax that varies with both weight and number of axles for heavy trucks, because spreading a given weight over more axles reduces road damage (Rufolo, Bronfman, & Kuhner, 2000); however, most states and the federal government raise road revenues from heavy vehicles through fuel taxes and registration fees. Raising revenue with efficient prices also serves to manage the use of the system. It is expected that efficient management would reduce the amount of road capacity required to meet any level of demand.

In the absence of better management of the road system, there are ongoing predictions of the need for massive investments. Although a pricing system would reduce the required investment as well as generate revenue, the large growth in demand for transportation over time indicates that more capacity will have to be added to the system. How that new capacity will be financed is a contentious issue. Although there is more consideration of pricing and tolling as finance mechanisms, they still represent a small percentage of the existing revenue sources.

Most revenue for the road system in the United States comes from fuel taxes and other charges to vehicle users, although there is substantial disagreement about whether road users pay the full cost of the system. Some of the disagreement comes from disagreement about what is a user cost. Fuel taxes are viewed as user charges by most, but some critics consider fuel tax revenue that goes to road construction and maintenance as a subsidy. More substantive disagreement occurs about items like property tax revenue used for local roads. Some view this as a subsidy for roads while others argue that local roads are primarily of use to local landowners. Other disagreements relate to how

general-purpose taxes on vehicles should be counted (Dean, 2003). Virtually all economists agree that vehicles should be charged for the externalities that they generate; however, there is disagreement about what the charge should be (Delucchi, 2007).

Cost allocation studies are done by both the federal government and a variety of states to determine whether different classes of vehicles are paying their proportionate shares of the cost of building and maintaining the road system. This is another area of substantial controversy, but the complexity of the issue precludes going into it in detail here.

Whether vehicles pay the full cost or not, there is agreement that the current method of funding the road system faces serious problems. The fuel tax has been a major source of road finance at both the state and federal levels; however, growing fuel efficiency and the prospect of alternative fuel vehicles raise questions about the adequacy of this source over time. In addition, the tax is typically set at a rate per gallon, so the purchasing power decreases with inflation, and there has been substantial resistance to increasing this tax. Because of the concerns about fuel taxes, there has been increased interest in more directly pricing the use of roads, either with a simple charge per mile (a vehicle miles traveled [VMT] tax) or some form of congestion pricing. An important drawback to these alternatives is the cost of collecting the revenue. As these costs decline over time, there is likely to be more extensive use of these alternative revenue sources.

One effect of improved ability to collect tolls or impose other prices is that it becomes more feasible for private firms to build roads, operate roads, or both. The concern with private firms operating roads is that the price that optimizes the use of the road may not coincide with the profit-maximizing price. Rather than having too much traffic because drivers are not paying the full cost of using a congested road, the road may be underused because of the high monopoly price.

The use of pricing has also raised questions regarding whether uniform pricing is the most efficient approach. HOV lanes are problematic. Their intent is to encourage carpooling, but they often appear to be underused or ineffective. When demand is relatively low compared to capacity, the total vehicle flow may be substantially reduced relative to general use of the lanes. A number of HOV lanes have been modified to allow non-HOV drivers to use them and pay a toll (high occupancy vehicle toll lanes are called HOT lanes). Effectively, solo drivers have the choice of congested but no-charge lanes or paying a fee for better conditions. This raises the question of whether such charging systems are more efficient than leaving all lanes unpriced and whether more than two prices might be efficient (Small & Yan, 2001). Because the value of saving time differs across people and for the same person under different circumstances, a better understanding of this distribution and of the effect of segregating lanes on traffic flow is needed. However, improvements in technology are likely to make complex pricing more feasible over time.

### Transit

As noted earlier, transit is often criticized for being costly and ineffective, with costs rising rapidly over time and transit's share of travel declining. The share trend can be affected by items like high prices for fuel, but increased ridership is likely to increase cost more than revenue and place further financial pressure on the system. There seems to be little likelihood of transit in the United States becoming self-sufficient. Hence, the major issue for transit is whether subsidies will keep up with rising costs or whether some changes in the way transit services are provided will improve efficiency (Klein et al., 1997). If congestion pricing is implemented, then demand for transit service will increase and reduced congestion will make bus service less costly and more reliable. Under these circumstances, the economics would argue for higher fares and lower subsidies, but much of the discussion related to making congestion pricing more acceptable to the public regards increasing transit service, as was done in London.

### Freight

Rapid growth in freight movement has placed strains on various parts of the freight system. The Federal Highway Administration (2007) identifies road congestion, intermodal transfer facilities, and capacity constraints on railroads as important issues. For trucks, the issues are very similar to the ones for automobiles in terms of congestion and financing the roads, although freight concerns are more concentrated on the need for improvements at specific bottlenecks.

Efficient pricing of roads would charge vehicles on the basis of their axle loadings and damage to roads. On the other hand, in the absence of such price incentives, regulations limiting weight per axle may also be inefficient. There is evidence that at least some truckers would be willing to pay for the extra damage if allowed to carry heavier loads (Rufolo et al., 2000). Expanding capacity for railroads that require additional construction would be at best a long-term solution given the high cost and other constraints on adding rail capacity. However, the railroads have shown significant ability to improve productivity since deregulation, and they may find other methods to address the capacity constraint.

## Conclusion

Transportation economics is a complex field that has received relatively little attention. However, the growing

divergence between demand for travel and the resources available to finance transportation infrastructure is focusing attention on both the methods of finance and the efficiency incentives. Transportation economists argue that more effective pricing of transportation would improve operation of the existing system while also providing funding for improvements. However, direct pricing faces technical, political, and public acceptance issues. Ongoing experiments and demonstration projects are likely to lead to gradual increases in the use of pricing, but a major shift does not appear likely in the near future. Hence, further increases in congestion on roads and financial pressure on transit systems seem inevitable.

## References and Further Readings


Abou-Zeid, M., Ben-Akiva, M., Tierney, K., Buckeye, K. R., & Buxbaum, J. N. (2008). Minnesota pay-as-you-drive pricing experiment. *Transportation Research Record, 2079,* 8–14.

Congressional Budget Office. (2008, January). *Effects of gasoline prices on driving behavior and vehicle markets.* Available at http://www.cbo.gov/ftpdocs/88xx/doc8893/01-14-GasolinePrices.pdf

Dean, T. B. (2003). Policy versus the market: Transportation's battleground. *Transportation Research Record, 1839,* 5–22.

Delucchi, M. (2007). Do motor vehicle users in the U.S. pay their way? *Transportation Research Part A, 41,* 983–1003.

de Palma, A., Kilani, M., & Lindsey, R. (2008). The merits of separating cars and trucks. *Journal of Urban Economics, 64,* 340–361.

Downs, A. (2004). *Still stuck in traffic: Coping with peak-hour traffic congestion.* Washington, DC: Brookings Institution.

Federal Highway Administration. (2007, January). *Financing freight improvements.* Available at http://ops.fhwa.dot.gov/freight/publications/freightfinancing/index.htm

Flyvbjerg, B., Skamris Holm, M., & Buhl, S. (2002). Underestimating costs in public works projects: Error or lie? *APA Journal, 68*(3), 279–295.

Forkenbrock, D. J. (2008). Policy options for varying mileage-based road user charges. *Transportation Research Record, 2079,* 29–36.

Gómez-Ibáñez, J.A., Tye, W. B., & Winston, C. (Eds.). (1999). *Essays in transportation economics and policy: A handbook in honor of John R. Meyer.* Washington, DC: Brookings Institution.

Holmgren, J. (2007). Meta-analysis of public transport demand. *Transportation Research Part A, 41,* 1021–1035.

Klein, D., Moore A. T., & Reja, B. (1997). *Curb rights: A foundation for free enterprise in urban transit.* Washington, DC: Brookings Institution.

Lam, T. C., & Small, K. A. (2001). The value of travel time and reliability: Measurement from a value pricing experiment. *Transportation Research Part E, 37,* 231–251.

Lave, C. (1994). It wasn't supposed to turn out like this: Federal subsidies and declining transit productivity. *Access, 5,* 21–25.

Leape, J. (2006). The London congestion charge. *Journal of Economic Perspectives, 20*(4), 157–176.

Mackie, P. (2005). The London congestion charge: A tentative economic appraisal. A comment on the paper by Prud'homme and Bocarejo. *Transport Policy, 12,* 288–290.

Meyer, J. R., Kain, J. F., & Wohl, M. (1965). *The urban transportation problem,* Cambridge, MA: Harvard University Press.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *American Economic Review, 62*(4), 591–604.

Parry, I. W. H., Walls, M., & Harrington, W. (2007). Automobile externalities and policies. *Journal of Economic Literature, 45,* 373–399.

Portney, P. R., Parry, I. W. H., Gruenspecht, H. K., & Harrington, W. (2003). The economics of fuel economy standards. *Journal of Economic Perspectives, 17*(4), 203–217.

Prud'homme, R., & Bocarejo, J. P. (2005). The London congestion charge: A tentative economic appraisal. *Transport Policy, 12,* 279–287.

Rufolo, A. M., & Bertini, R. L. (2003). Designing alternatives to state motor fuel taxes. *Transportation Quarterly, 57*(1), 33–46.

Rufolo, A. M., Bronfman, L., & Kuhner, E. (2000). Effect of Oregon's axle-weight-distance tax incentive. *Transportation Research Record, 1732,* 63–69.

Rufolo, A. M., & Kimpel, T. J. (2008). Responses to Oregon's experiment in road pricing. *Transportation Research Record, 2079,* 1–7.

Shoup, D. C. (2005). *The high cost of free parking.* Chicago: Planners Press, American Planning Association.

Small, K. A., Winston, C., & Yan, J. (2005). Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica, 73*(4), 1367–1382.

Small, K. A., & Yan, J. (2001). The value of "value pricing" of roads: Second-best pricing and product differentiation. *Journal of Urban Economics, 49,* 310–336.

Spielberg, F., & Shapiro, P. (2001). Slugs and bodysnatchers: Adventures in dynamic ridesharing. *TR News, 214,* 20–23.

Varaiya, P. (2005). What we've learned about highway congestion. *Access, 27,* 2–9.

Winston, C. (1998). U.S. industry adjustment to economic deregulation. *Journal of Economic Perspectives, 12,* 89–110.

Winston, C., & Shirley, C. (1998). *Alternate route.* Washington, DC: Brookings Institution.

# 65

# Urban Economics

James D. Burnell
*College of Wooster*

The field of urban economics was developed based on the observation that population and economic activity are concentrated in geographic space. Thus, one can define the field of urban economics as the study of the spatial relationships between individuals, households, and firms from an economic perspective. Much of urban economic analysis extends the maximizing behavior of individuals and firms from microeconomics to include how location affects this behavior. A focal point of urban economic analysis is how distance affects maximizing behavior. The fact that there is increased population and employment density at certain points in geographic space indicates that there are advantages of clustering of activity at certain locations. The concentration of activity also has consequences, both positive (i.e., enhanced productivity) and negative (i.e., increased congestion), and urban economists also consider these consequences.

Urban economics had its origins in the area of location theory and regional economics. The foundation was the role transport costs played in location decisions (Richardson, 1979), the hierarchy of trading areas, the system of cities (Christaller, 1933; Lösch, 1940), and the theory of land rent (von Thunen, 1966). The inclusion of geography in the decisions made by economic agents set the stage for urban economics to become a separate field in economics in the 1960s. A major reason was the important contributions of William Alonso (1964), Richard Muth (1969), and Edwin Mills (1967, 1972) in presenting theoretical and empirical analysis on the organization of urban space. The 1960s were also a period in which there was a significant focus on problems associated with the current spatial organization. Problems of poverty and social unrest experienced by large central cities made cities the focal point in the media, in public policy, and in academia.

The application of urban economic analysis has extended to subareas of the field, such as urban transportation, housing and real estate, and urban public finance. The field has evolved as the structure of urban areas has evolved. The focus of transport costs on location decisions has diminished and been replaced with the analysis of the impact of agglomeration economies on the size and composition of urban areas. The Alonso-Muth-Mills focus on monocentric urban areas has evolved to incorporate the growing multicentric nature of urban areas and the relatively rapid rate of urban sprawl. The empirical approach to analyzing urban economic phenomena has also evolved with the development of spatial econometric techniques that allow empirical modeling to account for the influences of the contiguity of urban space.

The remainder of this chapter is organized as follows. First, the concept of the urban area is considered from the economist's viewpoint. Next, a discussion of the elements that affect urban areas in an interurban context is presented. This section considers dominant explanations for the productivity of urban areas. The following section considers the intraurban relationships that exist and how they explain the urban spatial structure that exists in metropolitan areas. In particular, this section considers the monocentric model of urban land use as well as growing suburban and multicentric land use patterns. Finally, some outcomes and issues related to modern urban spatial structure are considered. The focus of this section is how urban economists address the spatial pattern of households.

## The Concept of an Urban Area From an Economist's View

Undergraduate urban economics textbooks typically provide different definitions of urban areas that reflect a hierarchy of urban areas, usually incorporating the definitions used by the Bureau of the Census. Currently, the Census defines an urban area as a community with a population of 2500 or more. A micropolitan area is an area with an urban core population of between 10,000 and 50,000 people, while a metropolitan area has an urban core of at least 50,000. The Census defines the spatial reach of these areas in terms of one or more counties that have a high degree of economic and social interaction with the urban core.

The urban core refers to an important city that is the focal point of the interactions with the areas outside the city. Cities are defined by political boundaries: the legal boundaries that define the political authority of these areas. Edwin Mills and Bruce Hamilton (1989) point out that to economists, political boundaries are less important than the market forces that contributed to the increased density of individuals and firms. The market forces that define the economic concept of an urban area can be considered from two perspectives: interurban and intraurban. These perspectives can be examined in terms of the location decisions made by households and firms.

One may consider interurban analysis as the study of competition across urban areas. Interurban analysis considers location decisions and their consequences for different urban areas. Households and firms evaluate the locational advantages of different urban locations. For example, a manufacturing firm may be assessing the productivity of the labor force for the type of workers it needs between the St. Louis and Minneapolis metropolitan areas. A household may evaluate which urban area to locate in based on the availability of employment, the cost of living, or the existence of desirable amenities. The economic concept of urban area in this case transcends the existence of the political boundaries of cities. The consequences of these decisions affect the growth and income-creation ability of the area.

Intraurban analysis addresses the location decisions and their consequences within urban areas. Once an urban area is chosen, both firms and households will then decide where within the area to locate. The market allocation of land among households and firms within an urban area was the focus of the Alonso-Muth-Mills monocentric model. Although the existence of political boundaries does not necessarily define the overall urban area in an economic sense, decentralization of households and employment within urban areas suggested that political boundaries do have some influence. The increasing fragmentation of urban areas since the middle of the twentieth century gave rise to a large number of suburban jurisdictions that could compete for households and firms based on their tax and service packages and their ability to use zoning to influence land market outcomes. The consequences of intraurban location decisions by households and firms and the existence of interjursidictional competition affect the fiscal viability of large central cities, the distribution of employment opportunity, and the distribution of income and minority groups throughout the urban area.

## Interurban Analysis: The Urban Hierarchy

The focus of interurban analysis is on the process of urbanization, in which economic activity concentrates at particular locations, and the factors that contribute to the extent of this concentration. The starting point of the urbanization process is location theory, whereby the profit-maximizing decisions of firms specifically considered how location affected profitability. Although location theory could explain the location decision of individual firms, urbanization also meant that there were other factors that would influence the size and growth of urban areas. Early explanations were based on central place theory, which described a hierarchy of urban places in a system of urban areas based on the market area for goods and services (Christaller, 1933; Lösch, 1940). The most recent emphasis on explaining the process of urbanization is on the importance of agglomeration economies, factors external to the firms that provide advantages of clustering economics activity (Fujita, Krugman, & Venables, 1999; Rosenthal & Strange, 2004).

Location theory includes transportation costs as well as the costs of inputs in the cost functions of firms. To illustrate the importance of transport costs in location decisions, the basic model assumed labor and capital costs were equal across space and the firm used a raw material input that was available at one location while the market for the firm's product occurred at a different location. Thus, the firm would have to choose whether to transport the raw material to the market location to produce its product or to locate at the raw material site and transport the product. The firm's profit-maximizing location would be that which minimized transport costs. The usefulness of this simple location decision was that it introduced the concept of an economic location weight that was determined not only by the physical weight but also by the unit transport cost per mile (O'Sullivan, 2009). Traditional examples include weight-gaining and weight-losing production processes. Weight-gaining processes include products that gain physical weight, such as water added in the beverage industries, or products whose transport costs are high because of their fragile nature. Weight-losing processes include mining and lumber.

Relaxing the assumption that the costs of other inputs used by the firm are equal across space allowed the spatial variation in cost and productivity of inputs to influence the location of the firm. Inputs prices that are generally considered to vary across space include labor, capital, land,

energy, and raw materials. Another factor considered important by firms is the impact of the tax and service package that can be offered by these governments as they compete among themselves for firms to locate within their boundaries.

Location theory could explain why firms choose among different locations, but it could not effectively address the fact that urban areas varied in size and in their ability to grow. Central place theory developed by Walter Christaller (1933) and August Lösch (1940) provided explanations for the location of market-oriented firms and the resulting hierarchy of urban areas that results.

The basic assumptions of the theory were that consumers were evenly distributed across space and transportation costs were equal in all directions. A firm existed at a particular location to serve the population. Consumers would travel to the firm to obtain the good, and the effective price paid by the consumer would be the price established by the firm plus transportation costs. The firm's market area would stretch to the point that the good's effective price was such that consumers were no longer willing to purchase the product from the firm. If the firm earned an economic profit, other firms producing the product would enter at another location.

As firms continued to enter, the market area of existing firms would decrease as consumers encountered lower effective prices due to closer proximity to firms. Firms would continue to enter only until normal profits were made and spatial equilibrium was reached. Each firm would have a spatial monopoly over a particular area, creating a spatial network for that product. Networks of different sizes for different goods would exist based on the size of the market area for the product. Smaller order goods are goods with small geographic market areas, and higher order goods have larger geographic market areas. A central place would be a location where one or more networks locate. The size of the urban area would be determined by the market area of the product with the largest geographic reach—the highest order good present at that location—and would contain all successively lower order goods. The market reach of its highest order good would define an urban area's position in the hierarchy. As an example, consider medical services. Small places would be expected to have a number of general practitioners and basic medical testing services. More specialized medical services would be found in a larger place. This establishes a hierarchical spatial link whereby residents of smaller places are linked to large places to obtain higher ordered services.

Central place theory's value in urban economics is primarily in its explanation of the pattern of retail activity. Other approaches to explaining the size and growth of urban areas also developed. Economic base theory posited that an area's economy was composed of two sectors. The basic sector was composed of firms that exported their product beyond the area's boundaries. The local sector was composed of firms that provided products to the area's residents. The growth of the area was determined by the export demand for the area's products. As income flowed into the area from exporting basic sector goods, a fraction of this income would be spent in the local sector through successive rounds of spending. The successive rounds of spending defined the multiplier effect—a dollar's worth of income would generate a larger amount of total income, depending on the size of the multiplier. The size of the multiplier was dependent on how much of the basic sector income was spent locally (McDonald & McMillen, 2007).

Economic base theory explains that the growth of an urban area is dependent on the change in the export demand for the area's basic sector products and how well developed the local sector is in providing goods and services for the area's population. This simple theory of urban growth has found wide practical application in the area of economic impact analysis, in which predictions are made regarding the impact of the attraction or departure of a key economic activity on the local area's economy.

The most recent emphasis in research on the process of urbanization has been on the advantages of the clustering of economic activity. This recent emphasis was stimulated by the emergence of the new economic geography based on the work of Paul Krugman (1991) in the area of international trade, but it has found much theoretical application in explaining the clustering of economic activity in urban areas (Fujita et al., 1999). Masahisa Fujita and Tomoya Mori (2005) describe the new economic geography as being a general equilibrium approach that specifically considers agglomeration in explaining the pattern and structure of urban areas. A primary consideration of the new economic geography is the specific incorporation of economies of scale and imperfect competition.

Scale economies are a fundamental requirement for increases in employment density in geographic space, since larger production facilities allow efficiencies that reduce average costs as output produced rises, and larger facilities lead to more workers locating in proximity to their jobs. The advantages of clustering are attributed to the existence of agglomeration economies, which are external benefits to firms that reduce their average costs at all levels of output.

Alfred Marshall (1920) originally identified three sources of agglomeration economies. One source is input sharing, in which final product firms purchase intermediate inputs from a specialized provider who is able to use economies of scale because of increased demand resulting from the clustering of the final product firms. Stuart Rosenthal and William Strange (2006) refer to input sharing as a form of local outsourcing. For example, computer hardware firms may locate near a computer chip manufacturer. The concentration of hardware firms allows the chip manufacturer to realize economies of scale that result in lower chip prices for the hardware firms.

A second source of agglomeration economies is labor pooling, which occurs when firms are able to draw from a large pool of specialized labor. The concentration of high-tech computer firms in Silicon Valley and biotech firms in the Boston metropolitan area reflect a large pool of highly educated individuals that offer small startup computer or biotech firms a labor force with requisite skills. This benefits not only firms who have a large demand for specialized labor, but also the workers who have other job opportunities if some of the small startups fail.

The third source is attributed to knowledge spillovers, whereby the presence and interaction of those with specialized knowledge about their products and production processes will stimulate a higher rate of innovation. The interaction of highly educated workers in Silicon Valley and Boston and the existence of (well-known) research universities increase the likelihood of innovation.

Agglomeration economies are classified according to the type of firm receiving the external benefit. Localization economies are external to the firm but internal to the industry, since they occur based on the extent of the presence of firms in the same industry. Urbanization economies are external to both the firm and industry and are attributed to the size of the area. Larger areas have a diverse set of firms and labor, which allows firms across industries to benefit.

The existence of agglomeration economies contributes to the size and growth of urban areas. Agglomeration economies increase the productivity of the area, which results in higher rates of growth. Increased demand for the area's products leads to higher labor demand. In areas realizing agglomeration economies, the existence of this greater productivity also provides a self-reinforcing effect. As production increases to meet the increased demand, there is an additional pull of firms who would benefit (O'Sullivan, 2009).

The consideration of the self-reinforcing effects of agglomeration suggests that agglomeration economies can be dynamic as well as static. Static agglomeration economies relate to the industrial and geographic dimensions of the effect of agglomeration economies on firms and the urban area and help explain why some urban areas are larger than others. The industrial dimension relates to the industries experiencing the benefits of localization or urbanization economies. The geographic dimension relates to the proximity of establishments in industries experiencing agglomeration economies and the diminishing effect of the agglomeration advantage as proximity between establishments decreases (Rosenthal & Strange, 2004). Static agglomeration economies refer to a one-time cost reduction associated with the clustering of firms (McDonald & McMillen, 2007).

Dynamic agglomeration economies relate to a time dimension of the impact and help explain not only the growth of particular urban areas but also the rate of growth. Dynamic localization economies cause continual reductions in the firm's average costs as the size of the industry in the area increases. Dynamic urbanization economies yield continual reduction in a firm's average costs as the size of the area increases compared to static agglomeration economies that are considered a one time reduction in cost (McDonald & McMillan, 2007). Rosenthal and Strange (2004) attribute dynamic agglomeration economies to knowledge spillovers whereby the acquisition and transfer of knowledge between firms occurs over time, resulting in cost advantages realized in the future.

Much empirical work has been done to determine the impact of agglomeration economies on the urbanization process. Rosenthal and Strange (2004) provide a comprehensive review of the empirical literature regarding agglomeration economies. The empirical literature follows a number of approaches to test various aspects of the impact of agglomeration economies on urban productivity. These include the testing for the importance of localization and urbanization economies, identifying the appropriate geographic level, and using a microbased versus an aggregate approach to estimating productivity effects. A brief representation of the empirical literature is considered here.

Edward Glaeser, Hedi Kallal, Jose Scheinkman, and Andrei Shleifer (1992) consider the role of knowledge spillovers on the growth of industry employment in the largest 170 cities. They test for three potential impacts of knowledge spillovers on industry growth. The first two relate to localization economies whereby within-industry knowledge spillovers lead to higher rates of growth, and the difference between them centers on the degree of local competition within the industries. On one hand, less competition allows innovating firms to realize the gains of the innovation internally and provides an incentive to innovate. On the other hand, greater local competition within the industry stimulates innovation as firms try to stay ahead of their competitors. The third potential impact of knowledge spillovers reflects urbanization economies. The more diverse the representation of industries, the greater the interchange of ideas across firms, which leads to faster growth.

The empirical model of Glaeser et al. (1992) considers the growth rate of employment in the six largest industries in 170 of the largest cities as a function of variables that measure the three potential impacts of knowledge spillovers, controlling for regional and natural characteristics that might affect local growth. Their findings show that growth is faster in cities where local industry competition is greater and where industry diversity is greater.

Rosenthal and Strange (2001) consider the three sources of localization economies to determine the level of geography at which the sources are important in explaining the concentration of industry employment. The authors hypothesize that the impact of the agglomeration advantages may be influenced by the spatial concentration of industry activity.

Some sources of agglomeration may be more important at close proximity, while others may be important over a larger area. They regress an index of spatial concentration for an industry on variables that measure input sharing, labor market pooling, and knowledge spillovers, controlling for transport cost and natural advantage. Versions of the model are estimated for the zip code, county, and state levels. The results provide evidence that labor market pooling is an important agglomerative source at all three levels of geography, input sharing was important at the state level but not lower levels of geography, and knowledge spillovers were important only at the zip code level.

J. Vernon Henderson (1986) addresses the importance of localization versus urbanization economies. From an industry location view, the difference is whether areas that specialize in particular industries have greater advantages or whether it is the size of the area that provides the greatest impact on productivity. Henderson employs a production function approach, where industry output is a function of inputs. He uses local industry employment as a measure of localization economies and urban size as a measure of urbanization economies. From his empirical results, Henderson concludes that there is strong evidence of localization economies for almost all industries considered and that the localization effects were large. He finds almost no evidence of urbanization economies. He also finds evidence that the agglomeration effects diminish for larger urban areas.

One of the empirical issues regarding estimating the effect of agglomeration economies has been the level of aggregation of the industry data. For example, Henderson (1986) used the two-digit level to define the industry. A more recent work by Henderson (2003) estimates plant-level production functions using panel data to provide a microlevel assessment of the effect of localization and urbanization economies and the ability to capture dynamic localization economies. Plant-level data is desirable since it is the plant that realizes the agglomeration economies at its particular location. Henderson considers the effect of localization economies and urbanization economies on what he identifies as machinery industries and high-tech industries. He also considers whether the plants are single plant or multiplant; single plants are more reliant on the local economic environment than plants that are affiliated with a corporation since the operation of these plants reflects the internal linkages determined by corporate decisions.

Henderson (2003) estimates plant-level production functions as a function of plant-level inputs plus measures to account for localization and urbanization economies. The number of plants in the same industry in the same county is used to measure localization economies and a measure of diversity of manufacturing employment for urbanization economies. Dynamic localization economies are measured by lagging the localization economies measure. The results indicate that localization economies are important for plants in high-tech industries but not for machine industries. The impact of localization economies is stronger for single plant versus multiplant affiliates. High-tech single plant firms also benefit from dynamic localization economies. There is no evidence that urbanization economies had an influence for either high-tech or machinery industries.

Rosenthal and Strange (2003) also use a microlevel approach with a focus on the birth of new establishments for six industries. They hypothesize that the presence of agglomeration economies results in new establishments clustering around existing establishments, while a dispersed pattern of new establishment locations would occur if there were no agglomeration economies. The sample used is composed of data measured at the zip code level, which is a departure from earlier analysis that typically used metropolitan areas as the geographic reference point. This allowed the authors to test for whether the effects of agglomeration economies diminish as distance from existing establishments increases. The authors focus on six industries that represent both innovative (e.g., software) and traditional (e.g., machinery) industries. The number of new establishments or the amount of new establishment employment in a zip code is regressed against variables that include the number of establishments per worker in the industry and outside the industry as a measure of competitiveness and diversity of economic activity.

Measures of urbanization and localization economies are also included. Urbanization economies are measured as employment outside the industry, while localization economies are measured as employment within the industry. To test whether the impact of the agglomeration economies diminishes with distance, Rosenthal and Strange (2003) include within- and outside-industry employment in a series of concentric rings around the initial zip code. The results are that localization economies are important for five of the six industries, while there is little evidence that urbanization economies matter. The results also indicate that the impact of localization economies diminishes with distance.

The review of interurban analysis indicates that many factors affect the location of economic activity across geographic space. The recent focus on the role of agglomeration economies is important given the changing nature of the economic base in many urban areas. As urban governments try to influence the location of firms to generate growth within their boundaries, the understanding of the type of agglomeration economies and the geographic extent of these economies become important information for policy makers.

## Intraurban Analysis: Urban Spatial Structure

Intraurban analysis considers how space is organized within an urban area. Historically, the development of urban areas was such that the intensity of land use was much greater at

the urban core, generally referred to as the central business district (CBD) and diminished as distance from the CBD increased. The urban land market is the mechanism by which land is allocated to the competing residential and business users. Residential and business users establish their willingness to bid for a certain location based on the location's value in terms of utility for residents or profit for business. Competition in the land market would allocate space to the highest bidder.

The origin of the land allocation process was the theory of land use developed by Johann von Thunen (1966). His contribution was to consider the competition for land use among users who valued proximity to a central location, and he developed the concept of the land rent function, whereby agricultural land users who desired to minimize the costs of producing and transporting their outputs to the market would bid for locations in close proximity to the market. Bids would decline for sites farther from the market center to account for the increase in transportation costs. A land rent existed based on the value that users place on the scarce locations surrounding the desired central location.

The modern version of the land allocation process is generally labeled the Alonso-Muth-Mills approach. Alonso (1964), Muth (1969), and Mills (1967) formally developed the theory of spatial equilibrium in the land market. The Alonso-Muth-Mills approach considers a competitive urban land market in which the demand for urban land is based on the utility-maximizing decisions of households and the profit-maximizing decisions of firms in a monocentric city. Firms value CBD locations since the CBD contains the transport node where firms export their products. A CBD location would reduce the cost of transporting products to the export node. For households, the CBD is the location of employment, and households value proximity to the CBD to reduce commuting costs. The value that firms and households place on proximity to the CBD defines their bid-rent function, which is the willingness of a user to pay for a location at a particular location from the CBD.

For firms, the bid-rent function is based on the profits of the firm:

$$\text{Profit} = pq - cq - txq - Rs, \tag{1}$$

where $p$ is product price, $c$ is unit cost, $q$ is quantity sold, $t$ is the transport cost per unit of distance, $s$ is the size of the site, and $R$ is the rent bid per unit of site size. $R$ is the amount the firm is willing to pay at distance $x$, holding the level of profit constant. Assuming competitive equilibrium where profits are zero and solving for $R$ yields the firm's rent bid function:

$$R = \frac{pq - cq - txq}{s}. \tag{2}$$

As distance from the CBD increases, the firm will reduce its bid by the increase in transport cost incurred for shipping its product to the CBD. The slope of the bid rent function,

$$-\frac{tq}{s}, \tag{3}$$

represents the reduction in land costs necessary to compensate for increased transport costs as distance from the CBD increases, and it measures the value of accessibility to the CBD. As compared with firms that have smoother bid-rent curves, firms with steeper bid-rent functions assign greater value to accessibility since the reduction in rent bid will be greater in order to compensate for the higher level of transport costs.

For households, the utility-maximizing decision defines the concept of the bid-rent function. Household utility depends on the consumption of housing services and other goods, and households face a budget constraint where income is spent on other goods, housing services, and commuting costs. Holding income and expenditures on other goods constant, consider the expenditures on housing services and commuting costs, which depend on distance to the CBD:

$$Rh + tx, \tag{4}$$

where $R$ is the rent bid per unit of housing services, $h$ is the amount of housing service, $t$ is the commuting cost per unit of distance, and $x$ is distance from the CBD. Spatial equilibrium requires that utility is constant at different distances. This means that expenditures on housing and commuting must remain constant for a constant level of utility. As distance from the CBD increases, households will change their bids to compensate for the change in commuting costs:

$$\Delta Rh + \mathrm{t}\Delta\mathrm{x} = 0. \tag{5}$$

The slope of the household's bid-rent function is

$$\frac{\Delta R}{\Delta x} = -\frac{t}{h}, \tag{6}$$

and it represents the amount rent bid will go down as distance increases.

Land market equilibrium requires that land goes to the highest bidder and that neither firms nor households can gain profits or utility by moving to a different location. The graphical representation of the urban land market is depicted in Figure 65.1. The slopes of the bid-rent curves reflect the value of accessibility to the three users. Commercial firms have the greatest value of accessibility to the CBD and outbid resident and manufacturing firms.

The Alonso-Mills-Muth model has been useful in explaining patterns of urban land use evident in American cities. The declining importance of the CBD and the increased presence of population and employment in the suburbs can be explained by a reduction in transport costs,

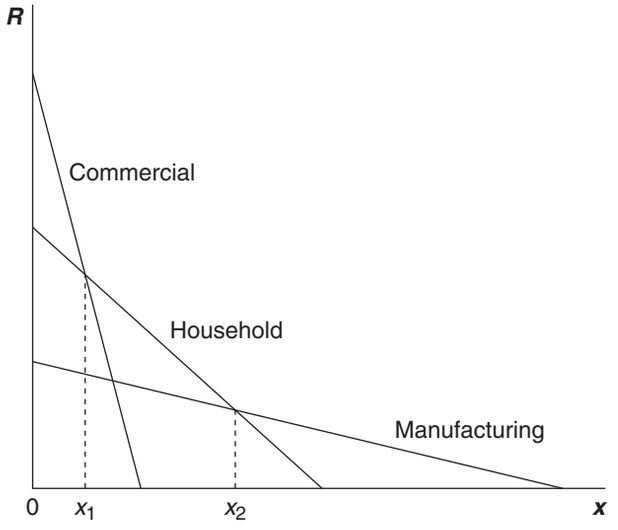


**Figure 65.1** The Urban Land Market

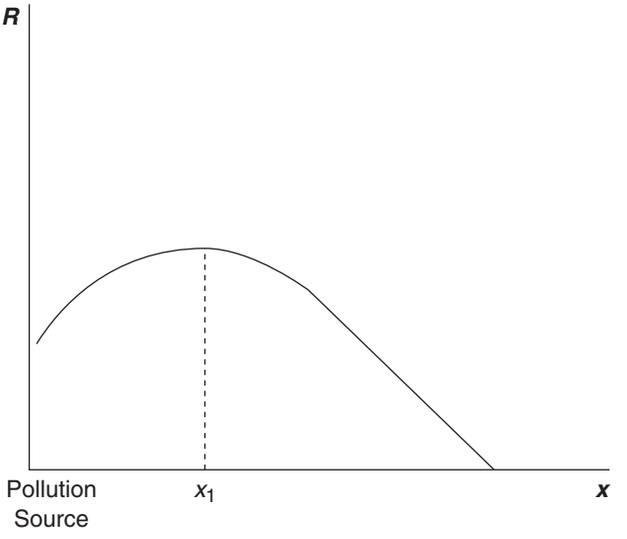


**Figure 65.2** Disamenities and the Household Bid-Rent Curve

which leads to flatter sloped bid-rent curves. The model also provided an explanation for the fact that lower income groups lived in the central city and higher income groups lived in the suburbs. Both Alonso (1964) and Muth (1969) consider this in their formulations of the monocentric model. The explanation is based on the income elasticity of demand for housing relative to the income elasticity of commuting. If the income elasticity of demand for housing were greater than the income elasticity of commuting, then high-income populations would locate in the suburbs where housing prices are lower and the amount of land available for housing is greater.

The Alonso-Mills-Muth model can be extended to consider other factors related to the characteristics of urban areas. Douglas Diamond and George Tolley (1982) incorporate the concept of amenities to determine their impact on households' bid-rent functions. Amenities are location-specific characteristics that yield utility to households but are not purchased directly. Diamond and Tolley incorporate amenities as an additional component of the household's utility function. The resulting utility-maximizing outcomes determine the values households place on amenities and yield bid-rent functions that can be positively as well as negatively sloped functions. As depicted in Figure 65.2, the positive-sloped portion of the function would represent proximity to a negative amenity, such as pollution, in which households would need to be compensated for locating at or near the source of the negative amenity. As the distance from the source increases, the effect of the negative amenity on the household diminishes, resulting in higher bids as distance from the negative amenity increases. This occurs up to $x_1$. Beyond $x_1$, the effect of pollution has dissipated, and compensating for commuting costs now dominates the bids made.

The Alonso-Mills-Muth model and its variations provided a theoretical structure for empirical work on urban spatial structure. Empirical approaches have included the estimation of population and employment density functions. Here, population and employment density are regressed against the distance from the CBD to determine the intensity of land use, with the monocentric model predicting that density decreases as distance from the CBD increases. Estimation of a hedonic housing price function is another empirical tool used to measure the value that households place on characteristics that comprise housing services. These characteristics typically include structural characteristics such as size of the dwelling and the number of rooms, quality characteristics such as the age of the dwelling, neighborhood characteristics such as crime rate or school quality, and proximity characteristics such as distance to the CBD or to particular amenities (Sirmans, MacPherson, & Zietz, 2005).

Early empirical work estimated population and employment density functions to test the predictions of the monocentric model of decreasing population and employment density as distance from the CBD increases. The early evidence provided by Muth (1969) and Mills (1972) supported these predictions. Even recent estimation of population density functions indicates population density is negatively related to distance from the CBD (McDonald & McMillen, 2007). Glaeser (2007) estimates the relationship between housing prices and distance from the Boston CBD and also finds support for the monocentric model.

Although the monocentric focus of the Alonso-Muth-Mills model remains the basis of much of the analysis of urban spatial structure, there has been recognition that the decentralization of population and particularly employment represents a transformation of the urban landscape (Glaeser & Kahn, 2003). Although the recent estimation of population functions and housing price gradients suggest that there is still the negative relationship between density and distance from the CBD, the empirical results suggest that the explanatory power of the monocentric model is relatively low.

Population decentralization has been attributed to factors such as poorer quality housing and the concentration

of minority groups and low-income groups in the central city that act to divert middle- and upper-income groups to the suburbs (Anas, Arnott, & Small, 1998). The politically fragmented nature of urban areas has also contributed to this decentralization through a Tiebout (1956) process, whereby households have the ability to match their preferences for local public services to particular communities.

Alternative explanations of the decentralization of employment have also been offered. In considering the intraurban location decisions of firms, Glaeser and Kahn (2003) offer three potential explanations of the draw to suburban locations. The first reflects the development of transportation infrastructure in the suburbs that offers savings in transportation costs compared to CBD locations. The second is that there may be differences in productivity between central locations and the suburbs. Knowledge and information spillovers are more likely to occur in the dense CBD areas, and this creates a productivity advantage for what Glaeser and Kahn refer to as idea firms—those that require high human capital and computers. The third explanation is that firms that do not benefit from these spillovers may find other cost advantages, such as proximity to workers, at suburban locations. They postulate that firms may follow the population to the suburbs. Based on their empirical research, Glaeser and Kahn conclude that the extent of decentralization in an urban area is greater when the presence of manufacturing firms in the area's industry mix is greater, that knowledge spillovers are important for idea-intensive industries, and that the labor force location has a strong influence on firm location.

The increasing focus on decentralization in urban areas suggests that there has been an evolution in the spatial structure such that many urban areas are considered polycentric. This relates to the observation that in many metropolitan areas, there exist edge cities (Garreau, 1991) that contain a concentration of office and commercial activity compared with other locations throughout the urban area. The existence of subcenters reflects agglomeration advantages away from the traditional CBD. In part, the technological advances of communication and the existence of the transportation infrastructure in suburban areas have contributed to this concentration of activity.

A major contributor to subcenter formation is also a contributor to the importance of monocentric city: agglomeration economies. Both localization and urbanization economies provide clustering advantages at locations away from the central business district. Robert Helsley and Arthur Sullivan (1991) and Denise DiPasquale and William Wheaton (1996) suggest that the advantages of these dispersed locations are enhanced by the increase in commuting costs associated with an increasingly dense monocentric city and by the existence of public infrastructure. Increased commuting costs necessitate higher wage costs for central business district firms to attract workers. Dispersed locations offer lower wages and land costs. The existence of public infrastructure at these dispersed locations provides an additional locational advantage for firms to cluster at particular locations. All of these factors enhance the external economies that occur as firms cluster at these sites.

Much work has been done to determine what empirical evidence exists for the existence of the polycentric form of urban areas. Three examples will be considered here. Kenneth Small and Shunfeng Song (1994) estimate polycentric employment and population density functions for the Los Angeles metropolitan area for the years 1970 and 1980. They find that the polycentric functions have greater explanatory power for the density patterns compared with the monocentric estimates. Daniel McMillen and John McDonald (1998) estimate employment density functions for Chicago for 1980 and 1990 to determine the effect of agglomeration economies on the existence of subcenters. They hypothesize that the presence of highway interchanges and commuter rail contributes to increased density at the subcenters. They also include proximity to subcenters as a measure of agglomeration economies. They find evidence that the agglomeration effects of shared transportation infrastructure and information and shopping externalities associated with subcenters contribute to higher employment density at these locations. Last, McMillen and Stefani Smith (2003) consider the number of subcenters in a metropolitan area. They consider the characteristics of 62 metropolitan areas and find that the number of subcenters increases as the population and commuting costs of an area increase.

## The Spatial Patterns of Residents

An important aspect of research in the urban economics field is the spatial pattern of the location of particular groups in urban areas and the consequences of this pattern. The spatial pattern that exists in most metropolitan areas is the centralization of low-income and minority groups in the central city or inner-ring suburbs. This spatial pattern evolves from household decisions to rent or own and from institutional factors such as lending and insurance practices and community land use policies.

Housing is important from an urban economics perspective because it is fixed in location and is a major component of housing services in the residential utility functions in the Alonso-Muth-Mills model. Along with a residential location decision, households must also decide whether to purchase or rent, which is referred to as *tenure choice.* Housing is also a durable good. The durability affects homeowners in two ways. First, housing deteriorates over time, and homeowners must decide on maintenance and repair. Second, homes generally appreciate in value over the length of ownership; thus, housing can represent an important asset to the owner. In the process of deciding to own a home, the household will require access to mortgage institutions. Housing market

outcomes have important implications for the distribution of population groups throughout the urban area. Different income and racial groups are affected by the availability of affordable housing, the preferences individuals have for neighborhood racial composition, and the existence of discriminatory behavior by institutions that facilitate home ownership.

The tenure choice decision of the household is based on the costs of homeownership versus renting. One element of the costs of home ownership is the mortgage payment. Mortgage payments are based on ability to obtain a loan and the resulting terms of the loan. The ability to obtain a loan is affected by the income and wealth of the household, the household's credit risk, and the value of the property. Lack of wealth has been a factor for low-income households generally; it has also affected the home ownership rates of African Americans (Charles & Hurst, 2002; Gyourko, Linneman, & Wachter, 1999; Wachter & Megbolugbe, 1992).

Racial segregation is a feature of the residential living patterns in metropolitan areas. The concern for racially segregated living patterns is the impact that segregation has on the outcome of minorities who live in more centralized neighborhoods. These outcomes include living in neighborhoods with poorer housing quality, lower quality educational opportunities, and restricted job opportunities (Cutler & Glaeser, 1997; Ross, 1998).

The preferences of racial subgroups are incorporated in residential location choice models by considering composition of a neighborhood as one of the neighborhood characteristics that affect the household's utility. In the case where a racial subgroup has an aversion to living with other racial subgroups, this will lead to segregated outcomes in housing markets. Analysis of racial preferences indicates that whites have the greatest aversion to living with nonwhites, particularly African Americans (Charles, 2005).

Another area that has been the focus of segregated living patterns has been the role of racial discrimination in housing and mortgage markets. In the case of housing markets, the analysis of how discrimination occurs centers on the behavior of real estate agents who provide information on housing and mortgage availability that may steer clients to neighborhoods of a particular racial composition (Yinger, 1995). The most recent empirical analysis to determine the existence and extent of racial discrimination in housing markets was sponsored by the Department of Housing and Urban Development in 2000 (Turner, Ross, Galster, & Yinger, 2002). Researchers used paired testing methodology, whereby equally qualified white and minority home seekers and renters interact with real estate and rental agents to determine whether they receive the same treatment in their housing search. Compared to the results of an identical study conducted in 1989, the evidence showed that African American and Hispanic home seekers and renters suffered from discrimination, but the incidence of discrimination was lower for African American home seekers and renters and Hispanic home seekers compared to 1989. Hispanic renters faced the same incidence of discrimination as they did in 1989.

Access to mortgage credit also plays an important role in the tenure decision. There is growing research interest in the question of how decisions by financial institutions affect urban residents. Much of the interest has centered on whether financial institutions provide access to mortgage credit based on the creditworthiness of the household rather than the household's membership in a particular income or racial group. Since the 1990s, a dual mortgage market has developed. The prime market is composed of low-risk borrowers who obtain loans with lower interest rates compared to borrowers in the subprime market. The subprime market is composed of those considered to be high-risk borrowers, and as a result, they pay higher mortgage interest rates than prime borrowers (Apgar & Calder, 2005). Mortgage discrimination occurs when borrowers are denied credit or are channeled to the higher rates in the subprime market or when the financial institution bases its decision by incorporating group membership as an additional determinant of the loan decision. The consequences are not restricted to the influence on the tenure decision of a minority or low-income household, but also include the increased risk of foreclosure.

The empirical research in mortgage lending discrimination typically shows that there are racial differences in mortgage lending outcomes, but it is more difficult to conclude that these differences are attributed to discrimination on the part of lenders. Alicia Munnell, Geoffrey Tootell, Lynn Browne, and James McEneaney (1996) provided a detailed analysis of factors that contribute to mortgage loan denials in Boston. Their empirical model includes variables that reflect the mortgage applicant's risk and cost of default, loan characteristics, and personal characteristics, with race as one of the personal characteristics. The results showed that African American and Hispanic applicants had a greater likelihood of being denied a mortgage loan than whites. Stephen Ross and John Yinger (2002) and Anthony Yezer (2006) discuss the criticisms of this approach, which range from exclusion of relevant variables that reflect creditworthiness to inability to account for the interaction of the applicant with the loan originator. The conclusion from these criticisms is that the results on the race variables will be biased.

A more recent approach applies the paired-testing methodology to the mortgage market. Ross, Margery Austin Turner, Erin Godfrey, and Robin Smith (2008) apply the paired-testing methodology to the pre-application process, whereby potential applicants interact with loan originators to obtain information regarding various types of loan products. The sample is composed of 250 paired tests in the Los Angeles and Chicago metropolitan areas. The results showed that African American and Hispanic testers in Chicago were less likely to be provided information, or they

received information on fewer products than white testers. The results for Los Angeles did not show a statistical difference in the treatment between African Americans, Hispanics, and whites. The implications for their results in Chicago are that minorities may have limited access to prime sources of mortgage credit, and this may lead qualified applicants to rely more heavily on the subprime market.

## Conclusion

As the presentation of this chapter demonstrates, the field of urban economics concerns itself with the location decisions of firms and households and how those decisions affect the population who live in densely populated urban areas. Urban economics addresses the circumstances and consequences of living in metropolitan areas. Its emergence as a separate field in the 1960s coincided with the changing urban spatial structure—the decentralization of households and employment and the centralization of poverty and minority households—and the consequences of these changes on central city and suburban residents.

Urban areas have evolved from being largely centers of manufacturing activity to having more service- and information-oriented activity as important components of their economic bases. The focus on the role of agglomeration economies has important policy implications for areas undergoing the restructuring of their economic base. Policy makers trying to retain or attract new firms need to be aware of the important contributions of agglomeration economies that exist in their areas.

Research in the field has also addressed the change in urban spatial structure that has occurred within metropolitan areas over the past 40 years. The theory of monocentric urban areas remains the foundation of understanding how the role of distance affects the location decisions of firms and households within urban areas. Extensions of the theory have been used to explain the development of multicentric urban areas and urban sprawl. Urban economics has also developed a number of subfields. Real estate economics and finance extend theoretical and empirical approaches to real estate markets. Governmental issues such as jurisdictional fragmentation and competition, land use policy, and urban fiscal issues are widely addressed. Policy makers interested in the future outcomes of their local areas need to be aware of the interaction that occurs between the pattern of increasing suburbanization and sprawl and its consequences not only for newly formed suburban communities but also for the central city, which may bear the burden of being home to an immobile population.

There are a number of areas that will be pursued in future research. As the national economy continues to undergo structural change, the impact of these changes on the productivity of urban areas will be explored. Much of the literature on agglomeration economies has focused on manufacturing. Future research should extend to the growing service sectors to determine the consequences of this restructuring on the size and viability of urban areas. The development of geographic information data by local governments and the development of spatial econometric software will become an increasingly important component of future research. These developments allow inclusion of spatial spillover effects and their influence on firm and household decisions. Finally, future research will also explore the consequences of urban sprawl and political fragmentation for metropolitan areas. An increasingly important area in urban policy is the concept of regionalism, where metropolitan governments may supersede particular decisions of local governments. Future research should develop a theoretical approach to consider cooperative decision making in metropolitan areas.

## References and Further Readings


Alonso, W. (1964). *Location and land use.* Cambridge, MA: Harvard University Press.

Anas, A., Arnott, R., & Small, K. A. (1998). Urban spatial structure. *Journal of Economic Literature, 36,* 1426–1464.

Apgar, W., & Calder, A. (2005). The dual mortgage market: The persistence of discrimination in mortgage lending. In X. Briggs (Ed.), *The geography of opportunity: Race and housing choice in metropolitan America* (pp. 101–126). Washington, DC: Brookings Institution.

Arnott, R. J., & McMillen, D. P. (Eds.). (2006). *A companion to urban economics.* Malden, MA: Blackwell.

Brueckner, J. (2001). Urban sprawl: Lessons from urban economics. In W. G. Gale & J. Rothenberg Pack (Eds.), *Brookings-Wharton papers on urban affairs 2001* (pp. 65–98). Washington, DC: Brookings Institution.

Brueckner, J. (2006). Strategic interaction among governments. In R. J. Arnott & D. P. McMillen (Eds.), *A companion to urban economics* (pp. 332–347). Malden, MA: Blackwell.

Carlino, G. A., Chatterjee, S., & Hunt, R. M. (2007). Urban density and the rate of invention. *Journal of Urban Economics, 61,* 389–419.

Charles, C. Z. (2005). Can we live together? Racial preferences and neighborhood outcomes. In X. Briggs (Ed.), *The geography of opportunity: Race and housing choice in metropolitan America* (pp. 45–80). Washington, DC: Brookings Institution.

Charles, K. K., & Hurst, E. (2002). The transition to homeownership and the black-white wealth gap. *Review of Economic Statistics, 84,* 281–297.

Christaller, W. (1933). *Central places in southern Germany.* London: Prentice Hall.

Cutler, D. M., & Glaeser, E. L. (1997). Are ghettos good or bad? *Quarterly Journal of Economics, 112,* 827–872.

Diamond, D. B., & Tolley, G. S. (1982). The economic roles of urban amenities. In D. B. Diamond & G. S. Tolley (Eds.), *The economics of urban amenities* (pp. 3–54). New York: Academic Press.

DiPasquale, D., & Wheaton, W. C. (1996). *Urban economics and real estate markets.* Englewood Cliffs, NJ: Prentice Hall.

Fischel, W. (2001). *The homevoter hypothesis: How home values influence local government taxation, school finance, and land-use policies.* Cambridge, MA: Harvard University Press.

Fujita, M., Krugman, P., & Venables, A. J. (1999). *The spatial economy: Cities, regions, and international trade.* Cambridge: MIT Press.

Fujita, M., & Mori, T. (2005). Frontiers of the new economic geography. *Papers in Regional Science, 84,* 377–405.

Garreau, J. (1991). *Edge city: Life on the new frontier.* New York: Doubleday.

Glaeser, E. L. (2007). *The economic approach to cities* (NBER Working Paper No. 13696). Cambridge, MA: National Bureau of Economic Research.

Glaeser, E. L., Hanushek, E. A., & Quigley, J. M. (2004). Opportunities, race, and urban location: The influence of John Kain. *Journal of Urban Economics, 56,* 70–79.

Glaeser, E. L., & Kahn, M. E. (2003). *Sprawl and urban growth* (NBER Working Paper No. 9733). Cambridge, MA: National Bureau of Economic Research.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., & Shleifer, A. (1992). Growth in cities. *Journal of Political Economy, 100,* 1126–1152.

Gyourko, J., Linneman, P., & Wachter, S. (1999). Analyzing the relationships among race, wealth, and home ownership in America. *Journal of Housing Economics, 8,* 63–89.

Helsley, R., & Sullivan, A. (1991). Urban subcenter formation. *Regional Science and Urban Economics, 21*(2), 255–275.

Henderson, J. V. (1974). The sizes and types of cities. *American Economic Review, 64,* 640–656.

Henderson, J. V. (1986). Efficiency of resource usage and city size. *Journal of Urban Economics, 19,* 47–70.

Henderson, J. V. (2003). Marshall's scale economies. *Journal of Urban Economics, 53,* 1–28.

Kain, J. (1968). Housing segregation, negro employment and metropolitan decentralization. *Quarterly Journal of Economics, 82,* 175–197.

Krugman, P. R. (1991). *Geography and trade.* Cambridge: MIT Press.

Lösch, A. (1940). *The economics of location.* New Haven, CT: Yale University Press.

Marshall, A. (1920). *Principles of economics.* London: Macmillan.

McDonald, J. F., & McMillen, D. P. (2007). *Urban economics and real estate: Theory and policy.* Malden, MA: Blackwell.

McMillen, D. P., & McDonald, J. F. (1998). Suburban subcenters and employment density in metropolitan Chicago. *Journal of Urban Economics, 43,* 157–180.

McMillen, D. P., & Smith, S. C. (2003). The number of subcenters in large urban areas. *Journal of Urban Economics, 53,* 321–338.

Mills, E. S. (1967). An aggregative model of resource allocation in a metropolitan area. *American Economic Review, 57,* 197–210.

Mills, E. S. (1972). *Studies in the structure of the urban economy.* Baltimore: Johns Hopkins University Press.

Mills, E. S., & Hamilton, B. W. (1989). *Urban economics* (4th ed.). Glenview, IL: Scott, Foresman.

Mills, E. S., & Lubuele, L. S. (1997). Inner cities. *Journal of Economic Literature, 35,* 727–756.

Munnell, A. H., Tootell, G. E. B., Browne, L. E., & McEneaney, J. (1996). Mortgage lending in Boston: Interpreting HMDA data. *American Economic Review, 86,* 25–53.

Muth, R. (1969). *Cities and housing.* Chicago: University of Chicago Press.

O'Sullivan, A. (2009). *Urban economics.* Boston: McGraw-Hill.

Richardson, H. W. (1979). *Regional economics.* Urbana: University of Illinois Press.

Rosenthal, S. S., & Strange, W. C. (2001). The determinants of agglomeration. *Journal of Urban Economics, 50,* 191–229.

Rosenthal, S. S., & Strange, W. C. (2003). Geography, industrial organization and agglomeration. *Review of Economics and Statistics, 85*, 377–393.

Rosenthal, S. S., & Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. In J. V. Henderson & J. F. Thisse (Eds.), *Handbook of urban and regional economics* (Vol. 4, pp. 2119–2171). Amsterdam: Elsevier North-Holland.

Rosenthal, S. S., & Strange, W. C. (2006). The micro-empirics of agglomeration economies. In R. J. Arnott & D. P. McMillen (Eds.), *A companion to urban economics* (pp. 7–23). Malden, MA: Blackwell.

Ross, S. L. (1998). Racial differences in residential and job mobility: Evidence concerning the spatial mismatch hypothesis. *Journal of Urban Economics, 43,* 112–135.

Ross, S. L., Turner, M. A., Godfrey, E., & Smith, R. R. (2008). Mortgage lending in Chicago and Los Angeles: A paired testing study of the pre-application process. *Journal of Urban Economics, 63,* 902–919.

Ross, S. L., & Yinger, J. (2002). *The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement.* Cambridge: MIT Press.

Sirmans, G., MacPherson, D. A., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature, 13,* 3–43.

Small, K. A., & Song, S. (1994). Population and employment densities: Structure and change. *Journal of Urban Economics, 36,* 292–313.

Tiebout, C. (1956). A pure theory of local expenditures. *Journal of Political Economy, 64,* 416–424.

Turner, M. A., & Ross, S. L. (2005). How racial discrimination affects the search for housing. In X. Briggs (Ed.), *The geography of opportunity: Race and housing choice in metropolitan America* (pp. 81–100). Washington, DC: Brookings Institution.

Turner, M. A., Ross, S. L., Galster, G. C., & Yinger, J. (2002). *Discrimination in metropolitan housing markets: National results from phase 1 of HDS 2000.* Washington, DC: Department of Housing and Urban Development.

von Thunen, J. H. (1966). *The isolated state* (P. Hall, Ed., & C. M. Wartenberg, Trans.). Oxford, NY: Pergamon Press. (Original work published 1826)

Wachter, S. M., & Megbolugbe, I. F. (1992). Racial and ethnic disparities in homeownership. *Housing Policy Debate, 3,* 333–370.

Yezer, A. M. (2006). Discrimination in mortgage lending. In R. J. Arnott & D. P. McMillen (Eds.), *A companion to urban economics* (pp. 197–210). Malden, MA: Blackwell.

Yinger, J. (1995). *Closed doors, opportunities lost: The continuing costs of housing discrimination.* New York: Russell Sage.

# 66

# Economics of Gambling

Douglas M. Walker
*College of Charleston*

Gambling refers to placing something of value at risk on the outcome of an uncertain event. Often, it is money being put at risk (in a bet), and the event can be anything: a flip of a coin, a football game, a roll of the dice, a poker hand, a lottery, or a horse or greyhound race. Games of chance can be divided into skilled and unskilled, the distinction being dependent on the extent to which random chance determines the outcome of the wager. In coin tosses, slot machines, and lotteries, for example, the outcomes are based entirely on random luck; the behavior of the bettor has no impact on the outcome. Games of skill, on the other hand, are those where the decisions and actions of the bettor can impact the results. These types of games would include poker and blackjack, for example.

There are a number of interesting policy and economic issues surrounding different forms of gambling, and this chapter introduces many of these, with a focus on gambling in the United States. Since legal gambling is a relatively recent phenomenon, especially in the case of casino gambling outside Nevada, readers can find many other interesting questions that have yet to be addressed by economic researchers. One reason that gambling represents an interesting case study is that it is almost always subject to strict government regulation. Throughout much of the twentieth century, most forms of gambling were banned by most state governments. Gambling has often been seen by many as an unsavory or immoral activity. The reversing of these bans is a relatively recent phenomenon, and with most states having some form of legalized gambling, the so-called moral opposition to gambling has largely subsided. The result is an enormous expansion in legalized gambling industries in the United States. Indeed, as of 2008, approximately 38 states had lotteries, 40 had horse racing, 17 had greyhound racing, 12 had commercial casinos, and 29 had tribal casinos (American Gaming Association, 2008; Walker & Jackson, 2008). The growth of legalized gambling, led recently by casinos, has increased significantly worldwide. The economic downturn that began in 2007, however, has hit the gambling industry very hard.

The most interesting gambling sectors for economists are lotteries and casinos, although Internet gambling has seen a dramatic increase in popularity over recent years, and recent legal roadblocks have brought this type of gambling into the spotlight. Although pari-mutuels (e.g., greyhound racing and horse racing) have a longer history in the United States, these industries are not as large as the casino industry and contribute much less to state governments than lotteries or casinos do, so this chapter does not focus on pari-mutuel betting.

Like other service industries, the gambling sector can be expected to have a variety of impacts on local and regional economies. In the next section, this chapter discusses theoretical issues surrounding the economic effects of gambling, as well as the limited empirical evidence that is available. Next is a discussion of policy issues that can be informed by economic research and suggestions for future research topics. The focus of this chapter is on casinos primarily and lotteries secondarily; this is because they are the highest volume industries and have been the focus of most of the economic research in the gambling literature. Readers who are interested in smaller gambling sectors, such as greyhound racing, will find that many of the issues discussed in this chapter are directly applicable to those other sectors. Overall, the emphasis in this chapter is outlining what is currently understood among gambling researchers. Since this is still a relatively young area in economic research, with only a handful of individuals focusing on gambling, there is still very little empirical evidence on the economics of gambling.

What makes the gambling sector unique is that it must be specifically legalized by state governments. Unlike other everyday goods and services that anyone is free to produce and sell, state governments determine the types of gambling that can be offered, the sizes of the venues, who may offer them, and the taxes that will be levied.

## Economic Impacts of Gambling

This chapter first addresses the benefits that are usually cited as reasons to adopt legal gambling. These effects include government revenues, which is the primary argument for lotteries and a major one for casinos; employment; consumer variety benefits; and complementary industry effects. Then this chapter discusses some of the potential negative impacts of legal gambling. These include regressivity of taxes, problem gambling, and a substitution effect with other industries. The discussion in this section includes an outline of theoretical issues, as well as a brief description of the available empirical evidence.

### Benefits

*Consumer Benefits*

As with other industries, one of the major benefits of the gambling sector is that it results in increased mutually beneficial transactions. That is, since both the buyer and seller of the product, in this case a lottery ticket or a casino game, receive benefits from their transaction, generally there is an increase in the overall well-being in society. Such transactions are the source of economic growth, since both parties are enriched by them. One need not emphasize the benefits to the seller (i.e., the casino owner or the state government offering the lottery), because the profits earned by the industries are quite visible, and the fact that the sellers earn profits is not unique to the gambling industry.

Some people find it surprising that the consumers also benefit from gambling, even if they do not win. To understand this, it is necessary to discuss in more detail the nature of consumer transactions in a free market. Gambling is similar to other goods and services in that something of value is exchanged for money. For example, when someone purchases a $3 box of cereal from the grocery store, the grocery store prefers the $3 to the box of cereal—if the posted price is $3, then it means the store would rather have the $3 than the box of cereal. The cereal cost the store less than $3, and the difference is the store's profit. On the consumer side of the transaction, the benefits are slightly more difficult to see. Generally, consumers' willingness to pay must be at least as high as the market price for the consumer to be willing to buy the product. In the case of the cereal, the shopper will buy the cereal only if he or she believes the cereal will yield at least $3 worth of benefits. For a consumer who really enjoys this particular brand of cereal, the expected benefit might be $10. In this case, the consumer receives a profit of $7, which is the difference between the value of the cereal to the consumer, or his or her willingness to pay ($10) and the price ($3). As shown in this example, both the buyer and seller receive a profit from the transaction. In economics, the seller's profit is referred to as *producers surplus,* while the benefit to the consumer is called *consumers surplus.*

Now turning to gambling as a service, one can see that transactions for gambling services are similar to that for the box of cereal. In the case of a lottery ticket, the price is $1. The state that sells the lottery ticket will, on average, return about 50¢ of each dollar to ticket buyers in the form of prizes and jackpots (Garrett, 2001). The remaining 50¢ is kept by the state to cover the costs of administering the lottery, with the remainder being kept as revenue for the state government to spend as it sees fit. (The costs section discusses lottery ticket revenues by the state in more detail.) The consumer pays $1 for the lottery ticket. The expected value of each ticket is around 50¢. That is, on average, the customer can expect to receive a 50% return on each dollar spent on the lottery. This is perhaps the worst bet of any legal form of gambling, in terms of the expected value. Aside from that, the odds against winning the jackpot are astronomical. Why, then, would anyone buy a lottery ticket? The reason is that the customers are not simply buying a return on their purchase prices. What are the benefits to lottery ticket buyers? Quite simply, enjoyment or entertainment. Many people find it entertaining and exciting to play the lottery. They enjoy the anticipation of seeing the winning numbers. They enjoy imagining what they would do if they won $200 million. Different people will value this experience differently.

Gambling at a casino is conceptually similar to the lottery. Almost all of the bets available in U.S. casinos have negative expected values. The casino makes its money by paying less than the true odds on winning bets. Typically the house edge is not that great, ranging from 1% to 15%, depending on the game. If a blackjack player bets and plays smart, the house edge is less than 5%. This means that for every $100 bet, the player should expect to lose less than $5. For slot machines, the house edge is higher, usually around 10%. As with lottery tickets, casino bets have negative expected value. The fact that millions of people go to casinos and play the games shows that they must receive some benefits from the experience. As with playing the lottery, casinos can be entertaining and fun. Gambling at a casino is also a more social experience than playing the lottery. The point here is many people enjoy the activity of gambling, and the benefits they receive from the activity (entertainment, excitement, etc.) apparently outweigh the price they pay (the house edge or the revenue from the lottery). This is similar to going to see a football game: Spectators expect to enjoy the experience enough to make it worth the price of the football ticket. Unfortunately, it is

extremely difficult to estimate the amount of consumer benefits from gambling.

It might seem unnecessary to take such effort to explain why gambling might be beneficial for consumers. Surprisingly enough, the consumer benefit of legalized gambling is rarely cited as one of the benefits to support the legalization or expansion of gambling. Yet there is little doubt that consumer benefits are the largest benefit to be gained from having legal gambling (Walker, 2007b, p. 622). There is any number of reasons that may explain this common oversight. It may be that since politicians legalize gambling with their sights on government revenue, they are not so concerned with consumer benefits. It may also be due in part to the fact that a small percentage of consumers develop a gambling problem (i.e., addiction). The problems faced by these individuals may overshadow the benefits that accrue to the vast majority of gamblers who do not have a problem with the activity.

Empirical evidence on the consumer benefits from gambling is astonishingly rare. Several studies have been performed in a few countries, such as Australia (Australian Productivity Commission, 1999) and the United Kingdom (Crane, 2006), but no really comprehensive empirical studies have been performed in the United States to date. Some researchers have at least been attempting to keep the consumer benefits issue in the debate over gambling, but this benefit often gets lost in all the talk about tax benefits, employment, and the social costs of gambling.

### *Government Revenues*

By far, the most obvious and commonly cited potential benefit from legalized gambling is revenue to the government. As discussed previously, the revenues from state lotteries can be substantial. Similarly, casinos can contribute a large amount of money to state government coffers. The government revenues from gambling are a strong argument for lotteries and casinos. However, the government revenue is one of the only arguments for the lottery. Casinos, as discussed in a subsequent section, may provide other economic benefits to a state that legalizes them.

Unfortunately, the issue of tax revenues it not as simple as it might first seem. Lottery revenues, for example, are often designated for supporting education. In states such as Georgia and South Carolina, lottery revenues are used to subsidize students' tuition. This additional funding for education may encourage legislators to designate less for education discretionary spending. That is, lottery-supported education may crowd out other education spending. The net impact of the lottery on education spending need not be positive. In fact, through lottery advertising, the state may imply that overall education spending has increased with the lottery, even if it has not. This would occur if nonlottery education expenditures were cut in an amount greater than the lottery-financed education spending. Unlike lotteries, tax revenues from legal casinos are not commonly designated for causes such as education. Such revenues are used to fund the oversight organization and sometimes for help for problem gamblers, with much of the tax revenue going into states' general funds.

Another interesting consideration for government revenues from legal gambling is that they represent voluntary taxes. That is, it is very easy for consumers to avoid paying these taxes; they can simply not buy lottery tickets or go to a casino. This argument is most commonly heard with the lottery, since nearly 50% of its sales represent tax revenues. One can argue that given government must raise revenue, consumers and taxpayers may prefer that the revenue be raised in ways in which it is easy to avoid paying the tax. Taxes on specific goods and services, like the lottery tax and taxes on casino gambling, fit this characteristic nicely.

As the states continue to face ever-worsening fiscal situations, they will continue to search for alternative ways of raising revenues. Raising more revenue using voluntary taxes is politically easier than cutting spending (benefits) or raising income taxes, property taxes, general sales taxes, or other unpopular taxes.

The total government revenues raised from gambling can be significant. As the American Gaming Association (AGA, 2008) reports, commercial casinos contribute a lot of money to the states, a total of $5.8 billion in the 12 states that had commercial casinos in 2007. Lotteries provide much more revenue. The North American Association of State and Provincial Lotteries (2009) reports that in fiscal year 2008, U.S. state lottery sales were about $60 billion, with net profits to the states around $18 billion. However, it is not clear that overall gambling revenue to the states more than offsets losses in other types of state revenues. The crowding-out issue has not been fully addressed by researchers. One recent empirical study on this topic suggests that gambling's net contribution—considering casinos, lotteries, greyhound racing, and horse racing—to states' budgets are relatively modest when they are positive (Walker & Jackson, in press). With respect to lotteries specifically, Thomas Garrett (2001) finds that state lotteries are designed to maximize the states' revenues from the lottery.

### *Employment Effects*

Another potential benefit of casinos is that they may have positive impacts on local labor markets. First, if a new business opens, it increases the demand for labor, which should push average wages higher. This benefits not only casino employees, but also other workers in the region surrounding the casino. Second, since a casino requires a major capital investment to build, the labor force required to build the facilities can be significant. Even once the building phase is completed, the casino will also need a significant workforce for its everyday operations. Often, a casino represents a major employer in its region. In this case, aside from simply putting upward pressure on wages and perhaps providing

jobs for the currently unemployed, the opening of a casino could increase the total number of jobs available.

Casinos may divert consumer spending away from other options, just as any new business is likely to do. But compared with many other businesses, casinos are rather labor intensive. That is, they tend to have more workers than other types of business. Consider a movie theater, for example. Given an equal number of customers, a casino would require many more employees than a movie theater would to operate effectively. The implication here is that even if there is a substitution effect whereby casinos divert spending and employment from other industries, it is certainly possible—and perhaps likely—that the casino will have a net positive impact on wages and employment. Unfortunately, the extent to which this occurs has not been addressed in much depth by economists.

Legalized gambling can have a significant impact on employment, but such cases will be limited. For lotteries and Internet gambling, for example, there is little or no employment effect. Horse and greyhound racing may have a modest employment effect but will probably not have as significant an impact as the average casino.

There have been very few econometric studies of how gambling affects employment and labor markets. Most of what has been written comes from the casino industry itself, and it amounts to a listing of employment data. The AGA (2008) provides a wealth of data and shows that commercial casinos do, in fact, hire a large number of employees. But as critics have suggested that some of those jobs may come at the expense of jobs in other industries, so the net effect of casinos on employment is unclear. One study that has addressed the issue rigorously is by Chad Cotti (2008). This paper examines casinos nationwide at a county level. Cotti finds that counties that introduce casinos tend to find increased employment is a result, but there is no measurable effect on average earnings. This is one of the first published studies to empirically address the employment effects of casinos.

### *Complementary Industry Benefits*

A final potential benefit from legalized gambling is the effects that the industry may have on complementary industries. As with employment, this issue is probably most relevant for casinos. The most successful casino model has proven to be the destination resort. Most casinos have an attached hotel, and patrons often stay, gamble, and dine in the same hotel-casino. Everything a vacationer needs is in one place. Despite the fact that often casinos can be resorts unto themselves, casinos can often do more business by agglomerating, or situating themselves near each other. This helps explain the great success of the Las Vegas strip. Many people go to Las Vegas instead of some other casino market because they know that there is a huge variety of casino resorts in Las Vegas. This variety provides important options for visitors. Even though competition often represents a negative for a particular business, in some cases, nearby competition can be beneficial.

Aside from any agglomeration effects that may benefit nearby casinos, a particular casino may also have a positive impact on other, noncasino industries. In Detroit, for example, there are three commercial casinos downtown. Casinos that draw tourists or even locals can provide business for the casinos, but visitors may also decide to patronize other area businesses. These businesses then may benefit from the casinos' existence. As discussed previously, the casino might have the opposite effect, substituting business away from other industries. Which effect is stronger is an empirical question that is going to vary by individual market conditions. As with other aspects of legalized gambling, there is scant empirical evidence on this issue. One recent paper to address this issue is by Cotti (2008). He finds that related industries see positive employment impacts and earnings spillovers from casinos. The other side of the coin (the substitution effect), however, has been studied more. Again, it is worth noting here that this issue applies mostly to casinos, less to racing, and not really at all to lotteries.

## Costs

### *Substitution Effect*

The previous discussion mentioned the possibility that casinos, in particular, may have complementary effects on neighboring businesses. On the other hand, casinos do act as a competitor to many types of business. In such cases, those industries will lose revenue and perhaps employment as the casino draws in customers. The state could in turn see a decrease in tax receipts. This substitution effect has been cited as a potential reason to avoid legalizing casinos, if the casino will cause more harm to other businesses than the benefits it creates. Consider an example in which casino revenues are taxed at the same rate that retail sales are taxed. Then, if casino spending were substituted dollar-for-dollar away from other industries, there would be no net tax effect from casinos. The same example could be applied to employment, so that the casino effect on employment would be neutral.

One might expect that as an entertainment industry, legalized gambling—whether the lottery, casinos, or horse racing—might impact different industries in different ways. This is fundamentally the case; in some cases, legalized gambling may act as a complement, and in others it is a substitute. It should be emphasized that the effects of a particular casino may be unique to that market. Simply because one sees a particular experience in one casino market does not mean that the same result would follow in another market.

Critics of casino gambling have long argued that gambling will not create new or better jobs because any jobs casinos create will be at the expense of other industries.

But it is not clear that this is the case. If employees actively seek casino jobs, it suggests that the casino job is the best opportunity available to them. However, if the casino causes other area businesses to close and those jobs disappear, then workers may have no option but to seek employment at the casino. Which case applies would depend on local economic conditions.

### *Regressive Tax*

Vertical equity is one of the basic principles of tax theory. It says that individuals with higher income levels can generally afford to pay more in taxes and that they should be expected to pay more for fairness reasons. For this reason, many people expect the government to raise tax revenues disproportionately for higher income individuals. Such a tax is called *progressive:* The proportion of taxes paid to income rises as a person's income level rises. However, if the proportion of tax paid to income rises as income falls, economists call this a *regressive* tax. Most people see such a tax as unfairly burdening the poor. (A tax for which the proportion of tax paid to income remains constant as income changes is called a *neutral* or *flat* tax.)

As an example, if a person with $10,000 income pays $1,000 in taxes, while a person earning $100,000 income pays $8,000, then this is a regressive tax: The poor person is paying 10% in taxes, while the rich person is paying only 8%. If the rich person were paying $10,000 in taxes, one would call this a flat tax, and if he or she were paying $12,000, for example, it would be a progressive tax. (Often *regressive* and *progressive* are terms applied to marginal tax rates, but this chapter is ignoring those for simplicity.)

Lotteries gained popularity with the states following New Hampshire's legalization of them back in 1964. Lotteries in the various states have raised an enormous amount of revenue for state governments. However, questions have arisen over who bears the burden of these taxes. If poor people buy a disproportionate share of the lottery tickets, then they will bear a disproportionate share of the taxes raised by the state-sponsored lottery.

One of the main areas of debate over lotteries is whether the lottery represents a regressive tax. Evidence has shown that poor people do, in fact, spend a larger share of their incomes on lottery tickets than wealthier individuals do. The lottery is sometimes referred to as a tax on people who are poor or bad at math. The first part—the poor—is due to the fact that poor individuals spend a disproportionate amount of their incomes on the lottery than relatively rich people. The second part—bad at math—is referring to the fact that the lottery is by far one of the worst legal bets available anywhere. Casino games, pari-mutuel betting, and even poker all carry higher expected values, in general, than playing the lottery. Thus, the lottery may prey on those who are unaware of the odds against them—people who are bad at math. The general consensus is that lotteries do represent a regressive tax. This issue has not been studied for casinos or other forms of gambling, however. The issue has surfaced for casinos, but it has not been analyzed. The regressivity issue has not really been a concern among researchers with respect to other forms of gambling.

One could argue that if state governments or voters were concerned about placing a disproportionate tax burden on poor people, this effect could be offset if the revenue raised were spent on helping those individuals who paid the tax. However, in many states, the lottery revenue is earmarked for subsidizing college students' educations. College students usually come from families with above average income. So even considering how the lottery revenue is spent, the lottery is generally believed to be regressive.

One issue to emphasize, however, is that the lottery tax is voluntary. That is, if poor people—or anyone else—would prefer not to pay the tax, they can simply not buy lottery tickets. According to this argument, then, the tax burden of lottery and casino taxes may not be a big concern. Of course, proponents of this argument might be sympathetic with the plight of the poor, and they might even believe that poor people would be better off if they spent their money on other goods and services, rather than on lottery tickets.

### *Problem Gambling*

Consumer theory in economics suggests that individuals are generally best off when they are sovereign and have freedom to choose how to spend their money. The economics of consumer behavior is based on individuals attempting to maximize their utility or benefits from consumption, subject to some monetary budget that they can spend on goods and services. Given that consumers see decreasing marginal utility from consumption, then their optimal or utility-maximizing bundle of goods and services is that for which $MU_a/P_a = MU_b/P_b = \ldots = MU_z/P_z$, for all the $z$ goods that they have the option of buying. Basically, this equation implies that consumers should spend each dollar on that good or service that will yield the most utility. Then when all the money is spent, they will have maximized utility. Theoretically, then, if the consumer has always chosen the best item to purchase, then the price-adjusted marginal utility for all items should be equal. Otherwise, the consumer could have purchased less of some goods (with relatively low *MU/P*) and more of others (with higher *MU/P*), resulting in a higher total utility. This behavior describes the typical consumer exhibiting rational behavior as it is taught in most intermediate microeconomics textbooks. In essence, the theory simply suggests that individuals make wise consumption choices based on prices and their estimates of potential benefits and costs from consumption. But perhaps gambling is different, or more precisely, perhaps some people behave differently when they gamble, compared with their behavior toward most other types of goods.

Psychologists, sociologists, and medical researchers have been studying gambling behavior, and over the past 20 years, the understanding of gambling problems has

advanced greatly. Just as some individuals may develop addictions to alcohol or drugs, the same can happen with gambling. That is, some people may become so-called problem gamblers. (There is actually a wide range of different severity of gambling problems, but the discussion here will not distinguish among them.) This behavioral disorder is similar to alcoholism or drug addiction. It is characterized by gambling too much—to such an extent that careers, relationships, and families are significantly harmed by the behavior or its effects. It is fairly easy to imagine how a person spending all of his or her time and money gambling would cause other problems in life. The American Psychiatric Association (1994) has estimated that between 1% and 3% of individuals develop gambling problems to varying degrees. As a result, these individuals sometimes engage in socially costly behavior. Some will commit crimes to get money for gambling. Others will default on debts, be less productive, or skip work. Problem gambling has even been blamed for breakups of marriages, bankruptcy, and suicide.

Estimating the costs of such effects of problem gambling is difficult at best. A first step is estimating how many people are affected by their gambling problems. Then the researcher must somehow estimate dollar values for the various negative impacts from problem gambling. This dollar value is multiplied by the estimated number of problem gamblers in order to arrive at a social cost estimate. Empirical estimates of the social costs of gambling are fraught with methodological problems, and there is an ongoing debate among researchers over how to deal with these issues (Walker, 2007b). Because of this ongoing debate and the uncertainty surrounding the social costs of gambling, it would be irresponsible to suggest that a specific cost estimate is correct. On the other hand, it is very simple to point to errors in the various social cost estimates that have been published or publicized. The youngness of this area of research is evident from the fact that social cost estimates have ranged anywhere from $800 to over $13,000 per problem gambler per year. These estimates have been derived, in many cases, from a variety of arbitrary assumptions. As a result, such empirical estimates must be taken with a grain of salt. At a state or national level, of course, if 1% of the population exhibits problem gambling behaviors, then the social costs of gambling can be significant. Simply because at this time the empirical estimates are of poor quality does not mean that the costs, unmeasurable as they seem to be, are not important and significant.

## Looking Forward: Policy Implications and Future Research

As is apparent from the discussion in the previous section, the research on the economic effects of gambling is still in the early stages of development. When empirical studies have been performed, their scope has usually been limited to small markets, to very short time periods, by unreliable data, or by some combination of these flaws. Therefore, it is very difficult to argue that there is some well-established empirical truth regarding how legalized gambling will affect a particular economy. At this time, the only thing one can say with certainty is that the economic effects of gambling will vary by location and region-specific characteristics. Obviously, much more research in this area is needed. Nevertheless, the political debates surrounding the economic effects of gambling have not waited on researchers to provide insight. Politicians and voters must deal with propositions to legalize or expand gambling or to change regulations. Often, the political debate is shaped by gambling proponents and opponents, with little support from economic research.

The policy implications with respect to lotteries are of limited importance at this time. Since most states already have some type of lottery, and for the most part, the states are extremely dependent on that money, there are few, if any, changes proposed in the status of lotteries. For example, it would be extremely surprising if any state that currently has a lottery were proposing to eliminate the lottery. Rarely will government simply cut off one of its revenue sources, especially during times of fiscal crisis. It is true, of course, that there is general agreement that the lottery represents a regressive tax. Even when the expenditures of the revenue are considered, the benefits of the lottery fall disproportionately on the relatively well-to-do, while the relatively poor pay a relatively large share of the burden. Although politicians and voters may be concerned with this type of redistribution of wealth, there is little chance that much will be done to change this.

Commercial casinos are a completely different story. During the past 20 years, these have been the source of intense debate in numerous states. As mentioned previously, 12 states currently have commercial casinos. Several others are actively working on legislation to allow them or have already had voted on casinos. For example, Kansas has approved casinos but is having trouble getting them started. Kentucky and Massachusetts have both already voted down casinos, but one can expect new proposals in the coming years until casinos are eventually approved.

Perhaps one should expect to see the most debate and the strongest push to adopt commercial casinos in those states that already have flourishing tribal casinos. This is because state governments receive limited revenues from tribal casinos. Since Native American nations are sovereign, their casinos are not governed by U.S. federal law or by state laws. However, tribes are required to sign compacts with their hosting states before tribal casinos can be offered. Often, such agreements include payments to the state for its agreeing to allow casino gambling. For example, the tribal casinos in Connecticut pay 25% of the slot machine revenue to the state's government. But the state could potentially do even better for itself if it were to allow commercial casinos too. Other states that have tribal casinos are going to be tempted to introduce commercial

casinos, which they would have the power to regulate and tax. This temptation is going to become even stronger as economic conditions around the country worsen. State governments are strapped for cash, and any opportunity to raise revenues will help politicians avoid even more difficult decisions.

## Cost-Benefit Analysis

When states have considered the introduction of casinos, a debate surrounding cost-benefit analysis typically ensues. One of the most common tools employed by state governments in evaluating the likely effects of introducing casinos is a cost-benefit analysis (CBA). These studies may be performed to a variety of degrees of technicality, but they almost always include the same components. The benefits and costs typically cited are those discussed earlier in this chapter. The exception is consumer benefits, which are rarely mentioned in such studies, and which are often ignored by politicians. The important benefits cited are usually tax revenues and employment. The costs focus almost entirely on social costs attributable to problem gamblers. These may include theft, bankruptcy, and decreased productivity on the job, as previously discussed.

The primary goal in CBA is to produce a concise summary of how legal casinos will affect a state's budget and economy. The simplicity of these studies—that their results can be summarized in one simple number, such as a net benefit of $10 million per year or a net cost of $10 million per year—makes them particularly attractive to politicians and even to voters. Policy makers may already know whether they wish to support casinos before seeing a CBA, but still the results of a CBA give them a piece of concrete data that they can cite to support their position. As discussed previously, however, in the context of social costs and consumer benefits, many of the costs and benefits associated with gambling are difficult, if not impossible, to measure. So even though politicians do rely on these types of studies to inform their decisions, they may be better advised not to, at least until this area of research improves significantly.

## Other Considerations

Although the debate over legalized gambling is often couched in terms of expected costs and benefits such as tax revenues, employment, and economic development, there are arguably more important, fundamental considerations that have been largely ignored in the debate. Obviously, it is important that voters and policy makers be aware of the potential consequences of their actions. So a good understanding of the potential costs and benefits of legalized gambling is critical. The uncertainty of monetary estimates—that some of the costs and benefits are inherently unmeasurable—necessarily makes any cost-benefit analysis arbitrary to some extent. Even if this were not the case, the more basic issues of consumer sovereignty, and the role of government in a free society should be acknowledged and contemplated.

As mentioned previously, economists generally assume that consumers make rational consumption decisions, acting in ways that they see as improving their welfare. Whether it is gambling at a casino or buying lottery tickets or football tickets, consumers expect the benefits from their consumption choices to outweigh the costs. But with gambling, there is the potential that some consumers may become addicted. The same is true of other goods and services, and numerous goods and services can be harmful if consumed in excess. Generally, in a free country, consumers are given the right to make their own choices, even if they may harm themselves. This is an important right in a free society. Yet with gambling and many other consumer goods, people's freedom is restricted. The extent to which consumer choice should be restricted is obviously debatable. But this fundamental issue receives little attention in the political debate over gambling.

Another important issue that should arguably be given additional consideration is the question of what the role of government in a free society should be. Should government restrict consumer choice for the good of consumers? Proponents of such a paternalistic role of government will point to drugs, alcohol, and other potentially dangerous goods as clear cases for which government regulation is necessary. On the other hand, staunch libertarians who view individual freedom as critically important will argue that there are many cases in which individuals may cause harm to themselves but that government cannot and should not attempt to protect everyone from their potentially bad decisions. The majority probably is more sympathetic with those who view regulation of gambling as justified. Still, a discussion of these issues should accompany policy debates. Unfortunately, gambling is seen by most as simply a tax issue.

## Directions for Future Research

Despite the growth of gambling industries worldwide, there is still relatively little economic research being performed. Of the different sectors of the gambling industry, lotteries have by far received most of the research attention. The focus has been primarily on the regressivity of the lottery tax, as well as on how the revenues are spent. There appears to be a general consensus in the literature that on net, the lottery is a mechanism that transfers wealth from lower income groups to higher income groups. Despite this undesirable outcome, governments are unlikely to abandon lotteries, since they are seen as providing easy money for the cash-strapped states.

The research on pari-mutuels has focused on how greyhound racing and horse racing are affected and how they affect other industries. There has been particular interest in the United States in the effect of allowing slot machines at racetracks. Racetrack owners have argued that allowing slots at tracks, so-called racinos, help keep the tracks competitive with the growing number of casinos. Overall, the

racing industry appears to be fairly stagnant. Still, some research is published on this industry regularly. But there are generally not many policy debates surrounding racing, other than whether to allow slots, so one should not expect much additional research on racing in the near future.

Internet poker has recently become the focus of significant attention since it was the target of recent legislation in the United States. Legal scholars can be expected to analyze what the effects of that law have been. As with lotteries, however, there are not many economic issues to debate with regard to Internet poker and online gambling. These appear to be mostly issues of consumer choice and regulation.

Casino gambling is by far the area in which most new research should be expected. Since casinos first spread in the United States beyond Nevada, data availability for testing the various economic effects of casino gambling has increased significantly. Among the issues that still need empirical research are the effects of casinos on state revenues, the effects of casinos on other gambling industries and other nongambling industries, the extent to which casinos in different states affect each others' revenues, the extent to which gambling availability contributes to problem gambling, the social costs of gambling (defining, measuring, and creating monetary estimates), and many more. In fact, almost any conceivable issue surrounding the economic effects of casinos would represent a significant contribution to the literature.

A handful of researchers dedicate much of their research effort on legalized gambling. As such work gets published and as policy makers look for more reliable evidence to support their policy decisions with respect to legalized gambling, one should expect more researchers to look at this fascinating industry.

## Conclusion

The economics of gambling is a wide-ranging subject, but most of the interesting issues under this topic are related to lotteries, horse and greyhound racing, and casinos. In terms of their revenues and contributions to state governments and the amount of academic and political debate they have inspired, casinos and lotteries are the most significant of the gambling sectors. This chapter focuses on those two industries, primarily.

Lotteries are important revenue sources for many states. Politicians like the lottery because it is a source of revenue, without which they might be forced to raise other taxes, cut spending, or some combination of both of those unpopular options. Despite the political attractiveness of lotteries, there has been some controversy in the economics literature over the extent to which lotteries raise their revenues at the expense of poor people. Empirical studies have confirmed that even considering how the states' lottery revenues are spent (e.g., many states use the revenues to subsidize college students' tuition), the lottery effectively transfers wealth from lower income to higher income individuals. Even considering this negative aspect of lotteries, there has been little push by voters or politicians to do away with lotteries. Of all the gambling sectors, economists' understanding of the effects of this one is the greatest. The lottery has been subject to numerous studies. There have not been too many new issues that have required the attention of researchers.

Casinos are receiving growing attention from researchers, but there is much work to be done. Casinos are the subject of intense political and academic debate, since many constituents have a strong interest in the outcomes of the debates. Among the issues that have been examined with respect to casinos are how casinos affect other industries within particular states, the effects casinos have on state-level employment and wages, and the net change in state revenues resulting from the introduction of casinos. Yet there are many issues for which empirical evidence simply does not exist. Since casinos began to spread across the United States in 1989, economists now have much more data on casinos and their effects. This will make empirical analysis possible. It is to be hoped that more researchers will become interested in the industry.

The literature on the economics of gambling focuses mainly on economic development effects of introducing casinos, for example. These benefits must be viewed alongside the potential social costs that may also come with gambling. Each jurisdiction's experience with legal gambling is likely different. Unfortunately, there has not been much empirical analysis of the economic effects of gambling. Even the relatively simple cost-benefit analyses that have been performed are potentially fraught with measurement errors. Gambling research is still young and is not very reliable yet. This suggests that researchers need to examine individual markets as well as more general relationships between gambling industries and other variables. There is much work to be done. The gambling industry is one that has been largely ignored by researchers. It is to be hoped that this is starting to change as the industry continues to grow.

## References and Further Readings

American Gaming Association. (2008). *State of the states: 2008.* Washington, DC: Author. Retrieved from http://www.americangaming.org/assets/files/aga_2008_sos.pdf

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Australian Productivity Commission. (1999). *Australia's gambling industries* (Report No. 10). Canberra, Australia: AusInfo.

Becker, G. S., & Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy, 96,* 675–700.

Clotfelter, C. T., & Cook, P. J. (1991). *Selling hope: State lotteries in America.* Cambridge, MA: Harvard University Press.

Collins, P. (2003). *Gambling and the public interest.* Westport, CT: Praeger.

Coryn, T., Fijnaut, C., & Littler, A. (Eds.). (2008). *Economic aspects of gambling regulation: EU and US perspectives.* Boston: Martinus Nijhoff.

Cotti, C. (2008). The effect of casinos on local labor markets: A county level analysis. *Journal of Gambling Business and Economics, 2,* 17–41.

Crane, Y. (2006). *New casinos in the United Kingdom: Costs, benefits and other considerations.* Unpublished doctoral dissertation, University of Salford—UK.

Eadington, W. R. (1999). The economics of casino gambling. *Journal of Economic Perspectives, 13,* 173–192.

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy, 61,* 279–304.

Garrett, T. (2001). The leviathan lottery? Testing the revenue maximization objective of state lotteries as evidence for leviathan. *Public Choice, 109,* 101–117.

Goodman, R. (1995). *The luck business.* New York: Free Press.

Grinols, E. L. (2004). *Gambling in America: Costs and benefits.* New York: Cambridge University Press.

Jackson, J. D., Saurman, D. S., & Shughart, W. F. (1994). Instant winners: Legal change in transition and the diffusion of state lotteries. *Public Choice, 80,* 245–263.

Kearney, M. S. (2005). State lotteries and consumer behavior. *Journal of Public Economics, 89,* 2269–2299.

McGowan, R. A. (2001). *Government and the transformation of the gaming industry.* Northampton, MA: Edward Elgar.

Morse, E. A., & Goss, E. P. (2007). *Governing fortune: Casino gambling in America.* Ann Arbor: University of Michigan Press.

National Gambling Impact Study Commission. (1999). *Final report.* Retrieved January 8, 2009, from http://govinfo.library.unt.edu/ngisc/index.html

National Opinion Research Center. (1999). *Gambling impact and behavior study: Report to the National Gambling Impact Study Commission.* Available at http://cloud9.norc.uchicago.edu/dlib/ngis.htm

National Research Council. (1999). *Pathological gambling: A critical review.* Washington, DC: National Academic Press.

North American Association of State and Provincial Lotteries. (2009). *Sales and profits.* Available at http://www.naspl.org

Smith, G., Hodgins, D., & Williams, R. (Eds.). (2007). *Research and measurement issues in gambling studies.* New York: Academic Press.

Thalheimer, R., & Ali, M. M. (1995). The demand for parimutuel horse race wagering and attendance. *Management Science, 41,* 129–143.

Walker, D. M. (2007a). *The economics of casino gambling.* New York: Springer.

Walker, D. M. (2007b). Problems in quantifying the social costs and benefits of gambling. *American Journal of Economics and Sociology, 66,* 609–645.

Walker, D. M., & Barnett, A. H. (1999). The social costs of gambling: An economic perspective. *Journal of Gambling Studies, 15,* 181–212.

Walker, D. M., & Jackson, J. D. (1998). New goods and economic growth: Evidence from legalized gambling. *Review of Regional Studies, 28,* 47–69.

Walker, D. M., & Jackson, J. D. (2008). Do U.S. gambling industries cannibalize each other? *Public Finance Review, 36,* 308–333.

Walker, D. M., & Jackson, J. D. (in press). The effect of legalized gambling on state government revenue. *Contemporary Economic Policy.*

Wynne, H. (Ed.). (2003). The socioeconomic impact of gambling: The Whistler Symposium. *Journal of Gambling Studies, 19*(2), 111–121.

# 67

# Economics of HIV and AIDS

Roy Love

*Health Economics Consultant*

The economics of HIV and AIDS is a strange creature. There is no so-called economics of mumps or economics of appendicitis. It is of course associated with sex, but then so are syphilis and gonorrhea, yet there is no economics of syphilis or even an economics of sexually transmitted infections. The difference lies in not only the fact that the virus is most often transmitted through sexual intercourse, an activity intrinsic to our humanity and therefore of universal interest, but also the fact that its presence is not always obvious, the consequences if untreated are fatal within only a few years, and there is as yet no known cure. It is therefore alarming in both its mode of transmission, striking at the very heart of one of our most intimate and pleasurable activities, and its catastrophic impact at the personal level. Because it affects adults who in many cases will be key household income earners, it also has profound social implications. Its impact on labor morbidity, productivity, medical and insurance costs, and public health expenditures affects business efficiency and a number of macroeconomic variables such as savings, labor and capital productivity, and private and public borrowing, and it raises important issues of the role of the state in prevention, care, and treatment. When the behavioral aspects of the risks of infection are included, then it is clear that it is a topic of considerable interest to economists.

To say that HIV is transmitted primarily through sexual intercourse is a reflection of its most common mode of transfer. More generally, the virus is transferred from one person to another by means of one person's bodily fluids entering another's bloodstream. It has a brief survival period out of the body and cannot be transferred via normal healthy skin contact, saliva, perspiration, or mosquito bites. The most common means are through heterosexual intercourse, homosexual intercourse, intravenous drug (IVD) ingestion by infected needles, and occasionally through contaminated blood products in a hospital or clinic environment. By far, the most common of these globally is the first, with a predominance in Africa, especially southern Africa where adult prevalence rates of above 20% are common, though IVD is an increasingly important source in former eastern bloc countries of Europe, central Asia, and India. In North America and western Europe, heterosexual, homosexual, and IVD have roughly equal weight as sources of infection but collectively amount to less than 1% of the relevant adult populations.

It is useful, before proceeding to a review of economic analysis in the area, to summarize the main characteristics of the disease. Only from 4 to 6 weeks following infection, and sometimes up to about 3 months later, can the presence of the HIV virus be detected. During this period, there will be no outward symptoms apart from a brief flulike illness, but the victim will be highly infectious. There then follows a period of 5 to 7 years when the HIV virus eats away at the infected person's immune system. As this process develops, the individual becomes gradually weaker, in due course becomes highly prone to opportunistic infections such as TB and pneumonia, and if left untreated, will generally die within about 10 years. In the context of the developing world, death tends to occur sooner because of poorer general nutrition and greater environmental health hazards. For biological reasons, women are more likely to be infected than men (that is, the transfer from an infected man into the vagina is more probable than the reverse). Morbidity and mortality thus tend to follow infection after a lag of some 5 to 10 years, making for a complex epidemiological cycle.

There is no cure, but a number of drug combinations are available that when taken regularly and continuously for life will raise and keep the individual's immune system at a level such that his or her life expectancy is considerably extended, though the individual will always remain HIV positive. Often, there are side effects that entail a shift to what are termed *second line* drugs, which tend to be more expensive. There is also around a 30% probability that a child born to an infected woman will be HIV positive. Treatment through intake of antiretroviral drugs can be costly, and there are important debates on international trade in pharmaceuticals concerning the rules of the WTO regarding which patent rights may be protected and when producers of generic supplies may legitimately trade. The chapter returns to this in a subsequent section.

For economists, all this gives rise to three broad areas of interest. One is the impact of HIV and AIDS on the economy at the macro, sectoral, or individual business level. The second concerns the choices made by individuals that expose them to the risk of infection and the consequences at household level, and the third is concerned with the whole area of public response, the role of the state, and the potential impact on public expenditure and taxation. A number of other issues arise from these, including the role of the social and institutional environments, the availability and cost of the drugs on which treatment depends, and the impact of stigma. The existing literature is composed of a very large number of heterogeneous papers, from which only a small representative selection is possible here. Although there are few classics specific to HIV and AIDS, most analysis draws on mainstream theory and its classic works.

## Impact on the Economy

This is a disease that principally affects sexually active adults, and to the extent that they are disabled by it, there is an impact on labor productivity and economic output. A simple list includes reduced productivity while at work through fatigue, increased absenteeism, higher than normal attrition rates and costs of recruitment, loss of skills, and time off to attend funerals, care for sick family members, and attend them in the hospital. Intergenerational effects will also appear as children in many poor communities are withdrawn from school and adult skills are not passed on (Bell, Devarajan, & Gersbach, 2004). At the macroeconomic level, these microeffects manifest themselves in reduced savings levels, reduced size of labor force (varying by sector and skill level), impact on public health expenditure, inflation due to increased business costs (group medical insurance, taxation, frequent recruitment), and possibly increased government borrowing, leading in turn to increased imports, balance of payments, and exchange rate problems. It is clear that in countries such as South Africa, Botswana, Zimbabwe, and Zambia, where HIV prevalence rates have exceeded 20% of adults between the ages 15 and 49 for most of the twenty-first century to date, the macroeconomic impact is likely to be considerable.

The most common ways in which attempts have been made to measure the macroeconomic impact of HIV and AIDS are either by cross-country econometric estimation or by application of macroeconomic models of varying degrees of complication (for reviews, see Haacker, 2004a, and Booysen, Geldenhuys, & Marinkov, 2003). Econometric estimation takes the form of including an HIV variable among others conventionally seen as affecting economic growth, such as savings rates, private investment, and education levels of the labor force. This approach, although able to produce statistically significant results (e.g., McDonald & Roberts, 2006), is less satisfactory as an explanatory or predictive tool than macroeconomic growth models that contain behavioral equations. The simplest of the latter start from the traditional textbook Cobb-Douglas type of production function, where output is a function of capital and labor (taking a variety of mathematical forms entailing different assumptions). These use aggregate data on capital stock (by value) and labor and insert assumptions on the impact of HIV and AIDS on these input factors to estimate the effect on productivity and output, thus having two basic growth scenarios: with HIV and AIDS and without. The basic model can be extended to include several skill levels of labor and, important in the case of many developing countries, the formal and informal labor markets. The following is a simplified version of an example from an application in Botswana, a country with one of the highest rates of HIV infection in the world:

$$Y = \gamma^{t} Es^{\beta s} Eu^{\beta u} K^{(1-\rho)},$$

where $Y$ is output, $Es$ and $Eu$ represent labor supplies of skilled and unskilled labor, respectively, and $K$ is the capital stock. The shares of output attributable to each factor are $\beta s$, $\beta u$, and $\rho = 1 - \beta s - \beta u$. An exogenous technological trend is represented by $\gamma^{t}$ (Econsult, 2006; Jefferis, Kinghorn, Siphambe, & Thurlow, 2008). The authors then explore the principal ways in which HIV and AIDS are likely to affect the labor supply and the capital stock and feed this into the model. The impact on labor supply will be affected by the degree of availability of antiretroviral treatment (ART), which requires additional assumptions. Other assumptions underlie the validity of such models in representing economic behavior. They assume, for instance, that the economy responds to changes in factor prices and that markets will clear, but they also usually assume constant returns to scale and a fixed rate of factor substitution. The authors of this study on Botswana concluded that the annual growth rate of GDP at market prices from 2001 to 2021 would be 4.5% in the absence of AIDS, 2.5% with AIDS, and 3.3% with AIDS plus ART (Econsult, 2006, p. 55, Table 5.3).

An interesting extension of this approach is where the concept of health capital is introduced as an additional capital variable. McDonald and Roberts (2006), for instance, incorporate (in an augmented Solow model) technological change and labor, plus physical, education, and health capital. Health capital itself is defined in a reduced-form equation as a function of lagged per capita income, education capital, nutritional status, HIV and AIDS prevalence, and proportion of the population at risk of malaria in a cross-country analysis. Proxies for health capital (the dependent variable) are life expectancy at birth and infant mortality rate. The results of the statistical analysis for the African sample indicated that a 1% increase in the HIV prevalence rate was related to a 0.59% decrease in income per capita. For the world sample, the decrease in income per capita was 0.5%, and for the developing world sample, it was 0.8%, each case having been brought up by a suspect high rate for Brazil.

An alternative means of estimating the impact of HIV and AIDS on the macroeconomy, which attempts to deal with the more complex and more realistic situation where the economy is broken down into a number of interacting sectors, emerged during the second part of the twentieth century in the form of computable general equilibrium (CGE) models for forecasting macroeconomic outcomes. As the name indicates, these models are (or claim to be) computable and hence testable versions of general equilibrium models of an economy. That is, they are versions of mathematical models in which the macroeconomy is the product of a number of behavioral decisions by consumers and producers at the microlevel of supply and demand in individual markets. The theoretical foundations of such models are found in the work of Walras, Arrow, Debreu, and others in the early and mid-twentieth century and are reflected in a substantial literature on the conditions that determine the possibility and existence of a general equilibrium in which demand equals supply across all markets freely and simultaneously. From this body of theory and the development of increasingly powerful computer capacity, economists working in economies where there is an abundance of current and historical data have been able to evolve ever more sophisticated computable models based on this theoretical foundation.

In practice, however, the degree to which many applications do in fact adequately recognize and incorporate the variety of experience at a microeconomic level has been questioned (Booysen et al., 2003; Johnston, 2008; Mitra-Khan, 2008). The focus in most applications of specific country forecasts of macroeconomic growth rates and associated variables (for example, by the IMF and World Bank) leads unavoidably to the primacy of macroeconomic and aggregated sectoral data sources, most frequently in the form of a social accounting matrix (a matrix representation of the national accounts of a nation, indicating the flow of activities from one sector to another). Even at this level, the data demands are considerable, and for many of the countries most affected by AIDS, the data are inadequate. Botswana is one of the better-off in this respect, and in the study referred to previously, the results of a CGE model with 26 productive sectors, 5 occupational categories, 3 regional areas, and a male–female breakdown are that the rate of growth of GDP from 2003 to 2021 would be 4.6% in the absence of AIDS, 3.0% with AIDS, and 3.4% with AIDS plus ART (Econsult, 2006, p. 101, Table 9.2).

Much of the work on applying CGE models to the macroeconomic impact of HIV and AIDS on an economy has taken place in South Africa, where the availability of data and local economic expertise, combined with levels of HIV prevalence above 20%, have stimulated much activity with the appearance of a number of such models. Some of these are demand-side driven, some are supply-side driven, and others have used a human capital approach. In a detailed review by Frederik le Roux Booysen et al. (2003) two of these (by ING Barings and the Bureau for Economic Research) are shown to forecast not only a difference in annual real growth of the South African GDP between an AIDS and no-AIDS scenario of –0.5 to –0.6 percentage points, but also a difference in predicted average annual growth in real per capita GDP of 0.9 percentage points in each model. The latter, in other words, is saying that real per capita growth in GDP is 0.9% higher in the presence of HIV and AIDS than without it. This seemingly perverse conclusion is created where the population growth rate is lower, as a result of HIV and AIDS, than growth in GDP. On the other hand, a CGE application to the Indian economy in 2006 concluded that the real GDP per capita growth rate between 2002 to 2003 and 2015 to 2016 was 6.13% with AIDS and 6.68% in the no-AIDS scenario. Real GDP itself was predicted to grow at 7.34% with AIDS, compared with 8.21% without AIDS, a difference of 0.87 percentage points (Ojha & Pradhan, 2006, Table 1). The latter is slightly higher than the corresponding figures for the growth rates with AIDS and without AIDS in South Africa, but too much should not be made of the differences since they will reflect different assumptions and specifications in the models and differences in the respective economies themselves.

Although these various models exhibit a high degree of mathematical sophistication, their output depends nevertheless on the quality of the data that is inputted. This includes the accuracy of existing measures of HIV prevalence (by which is usually meant the percentage rate of infection among adults aged between 15 and 49), which in most countries can be estimated from only a number of indicators since not all those infected will have come forward to be tested. In many developing countries, moreover, testing facilities are few and far between, and causes of death are often put down to an opportunistic disease such as TB or malaria. The most reliable figures historically have tended to come from testing of pregnant women at antenatal clinics, from which extrapolation, based on various assumptions, is made to the adult population as a

whole. The introduction of mobile testing equipment has enabled more accurate prevalence rates to be gathered through house-to-house surveys, but accurate measurement still remains a problem in many countries, especially if there is a recent history of civil disorder.

Such data uncertainty also makes it difficult to forecast the epidemiological progress of the disease and hence the likely impact on the labor force, especially when possible behavioral changes in response to public-awareness-raising campaigns are taken into account. Equally uncertain is the timing of the appearance of AIDS, which will depend on the degree to which ART is likely to be available in 10 to 20 years' time, its adherence rates, and the likely costs to the public health services. There is also in many countries a relative absence of reliable and relevant microeconomic information, such as the effect of HIV and AIDS on labor morbidity and productivity, for the purposes of the macroeconomic models. Will labor productivity be reduced by 20%, 30%, or even 50% by the HIV epidemic in certain countries? Assumptions very often have to be made on the basis of very little empirical evidence, and conclusions must be tested for their sensitivity to different assumed values.

The accuracy of the forecasts of such models on the macroeconomic impact of HIV and AIDS has also been questioned on the grounds that the division of labor in many countries is heavily genderized and that the impact at household level differs depending on whether an adult man or an adult woman (and in either case, a household head) is hit by AIDS. Evidence suggests that a higher proportion of female nonagricultural workers in sub-Saharan Africa are in the informal sector than the corresponding figure for men, and thus, that to the extent that women tend be more susceptible to HIV, the impact on the informal sector (which is substantial in many developing countries) will be understated by models that do not recognize the gendered segmentation of the labor market. On the other hand, where the burden of maintaining household production falls on women, their productivity is likely to increase, and hence, the negative impact will tend to be overstated (Johnston, 2008). This example illustrates the importance of understanding institutional and cultural constraints at microlevel.

Moreover, in addition to these obvious direct monetary costs of an epidemic, including both internal and external, there are welfare losses that are less easily measurable. For an individual infected by HIV who doesn't receive treatment, there will be not only a loss of income earning ability but also the loss of years of life and quality of remaining years. Nicholas Crafts and Markus Haacker (2004) illustrate this in a standard utility curve diagram in which expected lifetime utility is a function of annual income and life expectancy. The effect of HIV infection is to move the individual to a lower utility curve, at a point where both income and life expectancy are lower than before. Thus, the fall in income alone does not capture the total welfare loss to the infected individual. The authors then develop this model algebraically and use estimates of the value of a statistical life (VSL) to estimate the welfare effect of increased mortality across a sample of AIDS-affected countries. The VSL is a concept that measures the value of a variation in the risk of death and has been frequently estimated by surveys on individual willingness to pay for a given reduction in that risk. Given some qualifications regarding the paucity of data to estimate VSL in many developing countries, the authors show that the total welfare loss to countries such as South Africa, Zambia, and Botswana could range from 67% to 93% of each country's GDP (Crafts & Haacker, 2004, Table 6.2). Other theoretical work on welfare includes highly mathematical models (such as overlapping generations models) on the optimum control problem of social planners as they allocate limited resources in an economy to control an HIV epidemic at the cost of reduced levels of consumption (Shorish, 2007).

## Choices and Behavior of Individuals

As the discussion in the previous section indicated, any model of general equilibrium, or one that explicitly recognizes the links between consumers and producers at the microeconomic level and their aggregate impact in macroeconomic terms, must begin with assumptions about microlevel behavior. The common theoretical benchmark of a perfectly competitive equilibrium depends on a number of structural axioms, such as each consumer and producer having perfect knowledge, being a price taker, and striving to maximize either utility (satisfaction) or profits; diminishing marginal utility and diminishing marginal rates of substitution between goods demanded by the consumer and between factors of production used by the producer prevail (that is, indifference curves and isoquants are concave and production functions are convex); resources being perfectly mobile; and transaction costs being zero. Many if not most of these require modification to reflect real-life situations. The sort of rational individual they imply is traditionally referred to in the literature as *homo economicus*, or economic man, an expression criticized by many feminist economists as much for its conceptual roots as for its terminology.

The relevance of this to individual behavior in the context of HIV and AIDS may not be immediately obvious, but a moment's reflection shows that the sexual relationship is one in which each person concerned is making a choice between alternatives that have a number of possible outcomes. One set of outcomes in particular may affect the individual's health and hence future lifetime income. A common starting point for analysis at this microlevel is provided by human capital models in which individuals invest in education and health in order to enhance their future stock of health capital (and hence income earning opportunities) and the general quality of their own lives

and those of partners and near relatives. In doing so, they are forced to make choices between work and leisure, between activities that improve health or that have the potential to reduce it, and between more medical insurance and reduced current consumption and vice versa. These all lend themselves to traditional analysis of utility maximization, duly time discounted and adjusted for uncertainty, and many models of this type derive from original work by Michael Grossman (see Folland, Goodman, & Stano, 2007, for a summary). In the case of HIV, uncertainty takes the form of ignorance about a partner's HIV status, of the probability that they may be seropositive, and of the risk of becoming infected through a single act of intercourse. Unless the partner is in an advanced stage of AIDS, it is impossible to tell visually if he or she is infected, and hence, the exchange takes place in a context of incomplete information. In economic welfare terms, and somewhat unromantically and at its simplest, mutually agreed sexual intercourse involves an exchange of access to the most intimate parts of one's body in order to achieve an anticipated sensual satisfaction, whether or not money is present, and as such has the potential to be Pareto improving, except for the fact that complete information is absent in one or both of the parties to the exchange (Gaffeo, 2003).

Various mathematical models of microlevel behavior have been proposed and typically consider the numbers of people (or an individual) uninfected at the beginning of a period, the probability of becoming infected in each act of intercourse with an infected person, the number of partners during the period (though not always whether they are sequential or overlapping), and the probability that any one partner is infected. In one version, the trade-off then faced by an individual is the current benefit from more sexual activity and partners versus the costs involved if he or she becomes infected, where the infection is irreversible (Auld, 2003). In this example, the author predicts that where an epidemic is expected to get worse, certain individuals may actually increase their risky sexual activity, with the obvious policy implication that it may not be wise for governments to publicize the worst-case scenario as a means of reducing risky behavior. Another example constructs a lifetime model of utility from sex that takes account of the risks of infection and the relative pleasures from risky versus nonrisky sex, summed and discounted over an expected lifetime, from which the extent of the epidemic is determined significantly by the sensual difference between sex with or without condoms (corresponding to non-HIV risk and HIV risk) (Levy, 2002).

Underlying the development of such models is the interesting area of apparent dissonance between the predictions of rational decision making as conventionally understood in mainstream economic theory, as presumed in the human capital model, for instance, and the seemingly irrational behavior of individuals continuing to engage in HIV-risky sex despite knowing of its fatal consequences. It is unsatisfactory, however, to dismiss such deviance from the rational model of "economic man" as irrational and hence somehow outside the range of economic analysis. There are also parallels with persistent smokers, overeaters, and dangerous drivers, for all of whom knowledge of the outcomes of their actions is not sufficient in itself to bring about a change of behavior. In the case of an epidemic such as HIV where the status of a casual partner is unlikely to be known, plus the knowledge that the virus is not transmitted in every instance of intercourse, a person may be prepared to take the chance. There is some evidence that in the face of average figures on risk, many people are risk takers in that they feel that the risk to themselves is less than the average, leading to unrealistic optimism (Gold, 2006). Many—especially young people—have confidence that it won't happen to them. In such cases, a rational calculation is still being made. This is similar to the cognitive dissonance approach of George Akerlof and William Dickens (1982) in which choice is being made between beliefs such that a favored outcome can be justified.

It is also sometimes argued that in certain developing countries where life expectancy is already low because of general poverty, the future is heavily discounted in favor of short-term pleasures, especially if AIDS is known to not take effect for several years. This may also be associated with a fatalistic outlook. Evidence on the extent of this view is scarce, however. A factor often overlooked is alcohol consumption prior to sex. Social drinking of alcohol is a widely accepted phenomenon in many societies, including those in southern Africa where the HIV prevalence rate is exceedingly high. It is well known that among the immediate effects is confused thinking, but this does not imply totally random behavior. It does tend to mean that the gratification of immediate pleasures comes to the fore and that the future consequences of one's actions are again heavily discounted. Other possibilities suggest that an individual may decide in advance to avoid a risky activity, but when the opportunity arises or if the social context encourages it, the decision is reversed, in a form of situational rationality. In consumer theoretic terms, one thus has the phenomenon of preference reversal with implications for discounting models.

Some examples of indulgence in risky sex are entirely rational in the context of the individual at the time. Thus, many women are forced into transactional sex by economic circumstances and are in a weak position to negotiate safe sex. This may also be the case of many women in regular relationships including marriage. For others of either sex, the availability of ART may lead to an increase in risky sex, as may explain the positive correlation between the availability of ART and HIV incidence in the United States in the early 2000s (Lakdawalla, Sood, & Goldman, 2006). For others, the stage of the epidemic may be an influence: If it is believed that it will get worse in the future, then this may lead to an increase in the number of partners in the present for people so inclined. Alternatively, as the chance of any one partner being infected increases,

the marginal risk from an additional partner may be viewed as irrelevant.

What these examples all point to is that the challenge of changing behavior through public policy (such as campaigns for ABC: abstain, be faithful, condomize) is considerable, though there is evidence that condom use by sex workers has increased in many countries, sexually active people are having fewer partners in other countries, and young people (where they are free to do so) are delaying marriage and avoiding sex before marriage. This raises the question of the role of governments in prevention, care, and treatment of HIV and AIDS, discussed in the next section.

## Public Intervention

The usual arguments made for government intervention to limit the spread of the epidemic is that there is market failure in that incomplete and asymmetric information exists between the parties and that the externalities of spreading infection are not readily internalized. In theory, in the absence of this failure, the individual would be able to take his or her own informed action to avoid infection (that is, make rational choices that maximize his or her health capital), and any third party affected would be able either to negotiate compensation or to pay for the infection not to be transmitted (the theoretical context here is that of the Coase Theorem, which is explained in most intermediate microeconomics textbooks). A simple example of the latter is the cost of a condom, and an example of the former is the higher price charged by a commercial sex worker for sex without a condom. Information shortfalls refer not only to information about the nature of the disease but also to the HIV status of each partner, as known both of themselves and of the other. The consequences of this ignorance on wider society—that is, on the expansion of the epidemic and on the externalities of public health costs, taxation, economic productivity, and diversion of public expenditure from other social welfare priorities—are likely in the case of HIV and AIDS to be substantial, from which it follows that state intervention to preempt the growth of an epidemic and minimize the extent of related externalities is likely to be socially cost effective, though this still requires justification via cost effectiveness analysis of alternative strategies.

One approach is to encourage the internalization of the externality—that is, to eliminate it by having its probability contained in the transactional arrangement between the two people directly involved in an exchange. An example is through state subsidy in the provision of condoms, commonly known as social marketing. Other examples include the wide distribution of information about HIV and AIDS, including myths leading to stigma; establishment of national AIDS agencies; and increased availability and subsidy of voluntary counseling and testing. An interesting, if somewhat controversial, consequence of the last of these is that in a number of African societies, it is increasingly expected that couples intending to get married will undertake HIV tests (in subsidized clinics). These are examples of public intervention to internalize potential externalities. In general, such market-based compensatory negotiations, as theory suggests, between the two parties to an exchange will not be practical in instances of sexual trading, and in order to forestall or mitigate the potential impact of an epidemic on the wider community, the economy, and the public health services, there is a strong argument for state intervention.

The closest to a market solution tends to be found in treatment through private medical insurance, which in principle could cover physician fees, medication, hospital care, and loss of earnings but also contains the classic problems in the economics of insurance of adverse selection and moral hazard. The first refers to the tendency for those most likely to make a claim to seek to be members of an insurance scheme, and the second refers to the tendency of those within a scheme to maximize their claims income. In response to problems of this kind, insurance companies have adopted a variety of controlling regimes. In the case of HIV and AIDS, many require a statement of HIV status from new members and will charge a supplementary premium if the applicant is seropositive. This can have the effect, of course, of excluding those on low incomes, and in such cases, there are often publicly funded schemes that can step in. States in the United States provide HIV and AIDS treatment under the AIDS Drug Assistance Program (ADAP) for those otherwise unable to afford ARVs. To underline the point made about the social differences in the impact of HIV, it should be noted that in 2008, 59% of beneficiaries were African Americans and Hispanics and that only 35% were non-Hispanic whites (Kates, Penner, Carbaugh, Crutsinger-Perry, & Ginsburg, 2009), with an uneven geographic distribution. In most of the poorer developing countries, in which the highest rates of HIV prevalence are found, per capita disposable incomes are insufficient either to fund an adequate public health service through taxation or to purchase private health insurance. Only very few are able to afford private purchase of ART, which is often subsidized by international aid.

Market solutions are also constrained by social practices and institutions and tend to overlook relative positions of power. Of course, these can be taken as givens within which markets can be made as efficient as practically possible. However, it is not satisfactory, for instance, to take many traditional gender relationships simply as givens. Equalizing knowledge of HIV and AIDS does not equalize negotiating strengths between men and women when it comes to use of condoms or mutually agreed sex. In many societies, women also have fewer legal rights than men, fewer employment opportunities, and lower levels of education than men, all of

which reduce women's scope for economic independence and hence their market power relative to men. The same may be said of people with disabilities or certain minority ethnic groups. Care should be taken in assuming that vulnerability to HIV and AIDS is a function only of poverty or inequality. Botswana has one of the highest prevalence rates in the world and yet has one of the highest per capita GDPs in Africa, with universal primary education, high levels of literacy, and an extensive public health service. The trigger for the HIV pandemic appears to have been rapid social change associated with increased urbanization and rising social aspirations.

All such factors are present in decisions by governments to intervene to mitigate the spread and impact of HIV and AIDS. These may be national governments informing and supporting their own populations out of domestic taxation, but in many cases, they are also likely to be augmented by international aid from bilateral and multilateral donors whose concern is partly humanitarian and partly motivated by a broader agenda of civil stability, promotion of democracy, sustainable economic growth, and international security, in which the HIV and AIDS pandemic is seen as a threat. Governments also respond to private lobbying, often religious, which favors certain moral positions regarding sexual relations. In this field, economic calculus is only one factor among many that politicians have to consider when formulating policy.

## HIV and AIDS and Social Capital

Because of the way in which HIV affects working adults and, in many cases, heads of households, support within the immediate family is put under considerable strain, and help is often sought from members of extended families, neighbors, or the local community in general. Affected household members, in other words, draw on what has been termed their social capital for the additional resources needed to care for or treat an infected member or for help in holding the household together. Social capital is a concept that has been defined variously in the past and has been used loosely across a number of social science disciplines. A typical definition is found in the OECD Glossary of Statistical Terms (Organisation for Economic Co-operation and Development, n.d.), where it is described as "the norms and social relations embedded in the social structures of societies that enable people to co-ordinate action to achieve desired goals." It has also been defined as the "social structure which facilitates cooperative trade as an equilibrium" (Routledge & von Amsberg, 2003, p. 167) in which it is a means of reducing transaction costs in the absence of complete and enforceable contracts between individuals. As such, these definitions are also associated with a perception that social capital has an important positive role to play in the economic growth process, where in addition to the usual inputs of labor and physical (or monetized) capital, it becomes necessary to include norms of behavior, social networks, and trust as part of the informal institutional environment that supports most economic activities.

To the extent that HIV and AIDS damage social networks and trust within organizations and hence productivity, they also negatively affect economic growth (and are often implicit in the macroeconomic models discussed above). For example, additional work contracts may be required to cover absenteeism, or delays may appear in delivery of goods and services as informal personal links are weakened. In the HIV and AIDS context, social capital is also often brought into play at the microlevel, as indicated in the opening lines of the previous paragraph. A simple example of this would be an individual who helps a sick neighbor to purchase food, or it could be a group of neighbors who may help a struggling family with various farming tasks. However, the nature of HIV and AIDS is such that without ART, the pressures on the household are unremitting as time goes on, and the calls on the goodwill of extended family and neighbors eventually become resented. Another way of putting this is that the household's social capital is gradually used up. Furthermore, the stigma associated with AIDS can immediately restrict the amount of social capital available to a stricken individual or household.

It is, of course, extremely difficult to quantify social capital and hence to put a monetary value on it, and accordingly, it is common to use proxy indicators. An example is the measure of trust within communities, as gathered by the World Values Survey, for instance, which may then be taken as a dependent variable indicative of the level of social capital present in a range of countries. This can then be regressed statistically against assumed determinant independent variables including governance indexes, inequality data, and the prevalence rate of HIV and AIDS (David, 2007). The results show that a one standard deviation increase in HIV prevalence is associated with a 1% decline in the trust measurement index. And other things being equal, a country with a very high HIV prevalence rate would have a social capital level some 8% lower than one with a low rate (David). However, such studies are faced with serious methodological problems, largely deriving from the nebulousness of the concept of social capital and hence the characteristics of the dependent variable, plus problems of collinearity among the independent variables.

There are also a number of social norms, formally and informally institutionalized, that have a differentiating effect on the extent and type of social capital on which an individual can draw. In many developing countries, for instance, widows have fewer inheritance rights than men, and hence, their ability to survive independently is diminished, which may mean their reserves of social capital are both lower and used up more speedily. Orphans likewise

may find that their inheritance rights are lost through family manipulation.

## Economics of Antiretroviral Drugs

The availability of medication for treatment of HIV and AIDS has been a contentious area since the emergence of effective antiretroviral drugs that can control the immune deficiency created by the virus (though they cannot as yet eliminate it from the body). These drugs allow people infected by HIV to live active lives for many years, but they are demanding in terms of physician monitoring, side effects, and periodic regime changes caused by diminishing individual effectiveness. Initially, these drugs, usually now given in triple combinations, were very expensive but have come down substantially in price, especially for generic equivalents and, with international donor support, have enabled a gradual rollout in the most affected developing countries. There is no space here to go into the various regimes and their relative costs, but it is important to note that the regimes' provision and availability is part of the wider issue of drug supply and demand, in which several major economic problems are present.

The large pharmaceutical corporations that supply drugs tend also to be major researchers into the development of new drugs, but they argue that the opportunity to recover the costs of development would be compromised if competitors were able to reproduce the new drug too quickly. In more general terms, this is an issue of intellectual property rights. Since the nineteenth century, most industrialized countries have approached the problem by providing patent protection for a limited number of years, based on a patent registration scheme. Is this a case where the market has therefore failed? There are those who argue that the discoverer of a new drug has the advantage of being first in the field and that it is only because of effective political lobbying that patent laws have come into being, simply permitting excessive monopolistic profits to be made by beneficiary corporations. Evidence of this is difficult to come by and to interpret: For instance, profits higher than the industrial average will be defended as essential for the finance of research and development of new drugs. An alternative would be for more publicly funded research in universities or specialist research institutions. Underlying such proposals is the belief that medical research and its products are merit goods, having positive social externalities similar to the provision of literacy in the population, and that if left to individual choice, then the generated supply would be socially insufficient. Hence, a political decision is taken to support production either from taxation or, in the case of innovations, by patent laws. In the latter case, it will be the immediate consumer who provides the subsidy, which has some justice since the consumer is the direct user, but of course in social terms, this will tend to exclude the less well-off.

In most countries today, some form of patent protection exists, and it is relevant here to examine briefly how it is implemented and policed, especially in the context of a global economy and the ability of private sector companies in countries like India, Brazil, and China to reproduce, at considerably lower cost, many of the most advanced drugs developed by the larger, usually Western-based, pharmaceutical corporations. These nonpatent suppliers are generally referred to as *generic* producers in that they manufacture from a generic formula and sell products identical to the branded version of the original innovating company. In some cases, the original patent may have expired, but in others, this is not so. In the international market, this is clearly a threat to the ability of those companies that originally patented the drugs to receive the return they anticipated, and the problem becomes one of international trade regulation, in which the World Trade Organization (WTO) plays a central role. Following from the Uruguay Round trade talks (1986–1994) in which the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) emerged, the Doha Declaration on the TRIPS Agreement and Public Health was produced at the fourth ministerial meeting in 2001, in which the conditions were laid out under which a member country would be permitted short-term acquisition of generic drugs, usually in the context of a national emergency. The agreement made specific reference to HIV and AIDS, tuberculosis, malaria, and other epidemics, and to members' right "to promote access to medicines for all" (WTO, 2001, ¶ 4). This directive contained a condition that generic production (under what is termed a *compulsory license*) should be for domestic consumption only, but this clearly left the position of those countries without production facilities out of the picture. Consequently, the general council of the WTO amended the TRIPS agreement in December 2005 to waive the domestic consumption requirement so that (the UN-defined) least developed countries would be able to import drugs from generic suppliers abroad. This amendment contains a number of conditions to prevent cheaper generic products from being reexported to nonqualifying countries. To participate, of course, each member country needs to have in place an appropriate legal structure regarding patents and drug registration.

What this account of the complexities of international trade in medical drugs, including HIV antiretrovirals, illustrates is that it has evolved into a highly regulated market, both domestically within any country and internationally, and that any analysis has to pay attention to the considerable political lobbying power of transnational pharmaceutical corporations (*big pharma*, as critics tend to refer to them) and to the competitive nature of nation states in protecting their own citizens. Insights in this area are to be gained through analysis as much by political economy and institutional economics as by traditional theory. Nevertheless, the more successful the branded producers are in protecting their markets with higher prices, the more likely, given their limited resources, are many poor countries and their citizens to look to generic suppliers one way or another to meet their demands. In such a context, another characteristic of the market in health provision—namely, asymmetric information about complex products—allows charlatans and counterfeiters to operate in

the market, and there is evidence that the latter is occurring on a major international scale (see Web sites of the U.S. Food and Drug Administration and of the World Health Organization). In such an international environment, the ability of many developing countries to provide ART to all those infected by HIV is a major challenge that for the foreseeable future can be approached only with the support of international aid and technical assistance.

## Conclusion

In this necessarily short paper, it has been impossible to do justice to the very many applications of economic analysis to individual country studies of the impact, at national, sectoral, and household levels, of HIV and AIDS and of the cost effectiveness of various forms of treatment and care that have emerged over the last 20 years or so. This is a very rich field with many papers produced or sponsored by multilateral agencies, international NGOs, and private or academic research institutes, not all of which subsequently appear in journals. Limited space has also led to a concentration on HIV infection by means of heterosexual activity, since this is by far the most common means of transfer in those countries worst affected. The points covered, however, have an applicability to other means of acquiring infection, principally through homosexual activity and intravenous drug use, in which for the latter, at least, there are particular problems of adherence to ART, and in all cases, including heterosexual transmission, continuing problems of societal stigma.

## References and Further Readings


Akerlof, G. A., & Dickens, W. T. (1982). The economic consequences of cognitive dissonance. *American Economic Review, 72,* 3.

Auld, C. M. (2003). Choices, beliefs, and infectious disease dynamics. *Journal of Health Economics, 22,* 361–377.

Bell, C., Devarajan, S., & Gersbach, H. (2004). Thinking about the long-run economic costs of AIDS. In M. Haacker (Ed.), *The macroeconomics of HIV/AIDS* (pp. 96–133). Washington, DC: International Monetary Fund.

Booysen, F. le R., Geldenhuys, J. P., & Marinkov, M. (2003, September). *The impact of HIV/AIDS on the South African economy: A review of current evidence*. Paper presented at the TIPS/DPRU conference on the challenge of growth and poverty: The South African economy since democracy. Retrieved from http://www.tips.org.za/files/685.pdf

Canning, D. (2006). The economics of HIV/AIDS in low-income countries: The case for prevention. *Journal of Economic Perspectives, 20*(3), 121–142.

Crafts, N., & Haacker, M. (2004). Welfare implications of HIV/AIDS. In M. Haacker (Ed.), *The macroeconomics of HIV/AIDS* (pp. 182–197). Washington, DC: International Monetary Fund.

David, A. C. (2007, June). *HIV/AIDS and social capital in a cross-section of countries* (Policy Research Working Paper No. 4263). Washington, DC: World Bank.

Econsult. (2006). *The economic impact of HIV/AIDS in Botswana: Final report.* Available at http://www.unbotswana.org.bw

Folland, S., Goodman, A. C., & Stano, M. (2007). *The economics of health and health care* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Francis, A. M. (2008). The economics of sexuality: The effect of HIV/AIDS on homosexual behavior in the United States. *Journal of Health Economics, 27*, 675–689.

Gaffeo, E. (2003). The economics of HIV/AIDS: A survey. *Development Policy Review, 21*(1), 27–49.

Gold, R. S. (2006). Unrealistic optimism about becoming infected with HIV: Different causes in different populations. *International Journal of STD & AIDS, 17*, 196–199.

Haacker, M. (2004a). HIV/AIDS: The impact on the social fabric and the economy. In M. Haacker (Ed.), *The macroeconomics of HIV/AIDS* (pp. 41–94). Washington, DC: International Monetary Fund.

Haacker, M. (Ed.). (2004b). *The macroeconomics of HIV/AIDS.* Washington, DC: International Monetary Fund.

Jefferis, K., Kinghorn, A., Siphambe, H., & Thurlow, J. (2008). Macroeconomic and household-level impacts of HIV/AIDS in Botswana. *AIDS, 22*, S113–S119.

Johnston, D. (2008). Bias, not error: Assessments of the economic impact of HIV/AIDS using evidence from micro studies in sub-Saharan Africa. *Feminist Economics, 14*(4), 87–115.

Kates, J., Penner, M., Carbaugh, A., Crutsinger-Perry, B., & Ginsburg, B. (2009, April). *The AIDS drug assistance program: Findings from the National ADAP Monitoring Project annual survey*. Retrieved from http://www.kff.org/hivaids/upload/hiv040709pres.pdf

Lakdawalla, D., Sood, N., & Goldman, D. (2006, August). HIV breakthroughs and risky sexual behavior. *Quarterly Journal of Economics, 121*(3), 1063–1102.

Levy, A. (2002). A lifetime portfolio of risky and risk-free sexual behaviour and the prevalence of AIDS. *Journal of Health Economics, 21*(6), 993–1007.

Markos, A. R. (2005). Alcohol and sexual behaviour. *International Journal of STD & AIDS, 16*, 123–127.

McDonald, S., & Roberts, J. (2006). AIDS & economic growth. *Journal of Development Economics, 80*(1), 228–250.

Mitra-Kahn, B. H. (2008). *Debunking the myths of computable general equilibrium models* (SCEPA Working Paper No. 2008-1). New York: Bernard Schwartz Center for Economic Policy Analysis.

Ojha, V. P., & Pradhan, B. K. (2006). *The macro-economic and sectoral impacts of HIV and AIDS in India: A CGE study*. New Delhi, India: United Nations Development Programme.

Organisation for Economic Co-operation and Development. (n.d.). *Glossary of statistical terms.* Retrieved from http://stats.oecd.org/glossary/search.asp

Routledge, B. R., & von Amsberg, J. (2003). Social capital and growth. *Journal of Monetary Economics, 50*(1), 167–193.

Shorish, J. (2007). *Welfare analysis of HIV/AIDS: Formulating and computing a continuous time overlapping generations policy model* (School of Economics Discussion Paper No. 0709). Manchester, UK: University of Manchester.

World Trade Organization. (2001, November). *Declaration on the TRIPS agreement and public health.* Retrieved from http://www.wto.org/english/thewto_e/minist_e/min01_e/mindecl_trips_e.pdf

# 68

# Economics of Migration

Anita Alves Pena and Steven J. Shulman
*Colorado State University*

The economics of migration is a sizable topic area within economics that encompasses broadly defined studies of the movement of people within and across economies. Studies of intranational, or internal, migration focus on movements within a country's borders, whereas studies of international migration (emigration and immigration) focus on movements across international boundaries. The economics of migration spans several subdisciplines within economics. Both microeconomists and macroeconomists are interested in how migration affects markets for labor, other factors of production, and output. Labor economists are particularly interested in migration as it is an important determinant of labor market outcomes such as wages and employment. Public economists and public policy makers are interested in the effects of migration on the social surplus and in the interrelationships between migration and public policy instruments. Economists studying economic development and international economics are interested in how migrations affect economic outcomes in the developing world and in the global economy broadly.

This chapter outlines the key theoretical elements of microeconomic and macroeconomic models of migration and presents empirical evidence from representative applied studies of intranational and international migration to date. Special attention is given to the debate about immigration into the United States and its consequences for both natives and immigrants.

## Microeconomic Theory

### Economic Benefits and Costs

In his classic article, "A Pure Theory of Local Expenditures," Charles Tiebout (1956) hypothesized that people "vote with their feet" by migrating to localities with public expenditure characteristics that best fit their personal preferences. Tiebout's model illustrates, for example, how residential decisions are related to taxation and expenditure characteristics such as local tax rates and the quality and quantity of publicly provided goods such as education and local amenities.

In microeconomic models of migration, "voting with one's feet" is generally modeled by assuming that individuals make decisions regarding remaining at a current location versus moving to a preferable location. In its simplest form, the model may be described by agents maximizing net present value of lifetime earnings and engaging in migration if the difference in lifetime earnings between a potential destination and the agent's origin is positive and greater than migration costs. Thus, agents make investments in their human capital by moving to where their economic opportunities, as measured by lifetime earnings net of migration costs, are improved (Sjaastad, 1962).

Agents in these migration models are assumed to maximize their welfare by comparing net economic benefits (benefits minus costs) at an origin and at alternative locations in a large set of potential destinations. The decision therefore is not only whether to migrate but also where to migrate if a migration is to be undertaken. Economic benefits and costs are not the same thing as financial or accounting benefits and costs. Instead, economists use surplus areas (e.g., consumer and producer surplus) to define these concepts. In addition to accounting costs, opportunity costs are included to form the cost definition, and welfare is measured relative to some status quo.

In extensions to the basic model, agents take into account factors that influence economic and psychic benefits and

costs such as labor market variations, public policy and environmental attributes of various locations, and personal characteristics and circumstance. Furthermore, expected benefits and costs may differ depending on whether an individual will be migrating with or without family, legally or illegally, and so on. To complicate this further, precise values of economic benefits and costs often are unknown, and thus agents are thought to make decisions based on *expected* net benefits, instead of deterministic ones. Expected values take into account probabilities of uncertain outcomes.

If agents are assumed to be utility maximizers, then agents maximize expected utility by choosing to migrate to a destination from their set of potential destinations (which includes a stay at origin option) that maximizes expected net benefits where expected incomes and costs are mapped into utility terms. Expected net benefit to a person from making a migration is the difference between that agent's expected utility at the destination and his or her expected utility at the origin plus expected migration costs.

On the benefit side, expected income/utility may include both expected wage earnings and expected supplementary nonwage income such as public aid payments. More broadly, economists may include value of factors such as participation in public schooling and environmental amenities. Since both wage and nonwage income are expected measures, the probability of employment (and likewise the probability of receiving aid) should be included in the calculation. Agents compare these expected benefit values for each potential destination to expected utility at the origin, where again this may depend on probabilities of employment and of nonwage income as well as differences in generosities.

On the cost side, expected costs may be thought to be a function of monetary, opportunity, and psychological costs. An agent's total expected monetary cost of migration includes direct travel expenses. Opportunity costs of migrating to the United States include any foregone income at the origin and account for travel time and distance. Expected psychological costs associated with migration may include elements such as leaving family or one's homeland and may depend on travel distance and time. If a migrant intends to return to the sending location, then expected costs should represent round-trip costs. In the case of illegal migration, expected costs may include probabilities of apprehension and deportation and related costs (e.g., court costs, opportunity costs of time, and additional psychological costs) and any monetary payments to agents such as border smugglers for assistance in the trip.

All values on both benefit and cost sides may depend on a particular time of migration given varying political and economic contexts. Several locational attributes should also influence the propensity to choose one destination over another. High unemployment rates and other negative indicators of labor market conditions, for example, should be associated with decreases in the probabilities of employment at the destination and origin. Increases in average wages of similar workers and potential values received from social service programs, hospitals, and educational systems should be positively related to expected incomes. For those migrating illegally, border patrol intensity and the political economy of immigration policy may affect benefits and costs of migration. Furthermore, personal and professional networks may increase the probability that one crosses successfully and of employment at a destination. Networks also may increase the probability of receiving public aid benefits if experienced friends and family members help in the application process and may decrease both the monetary and unobserved psychological costs of crossing.

These considerations can shed light on applied questions such as the determinants of illegal immigration from Mexico into the United States. While the magnitude of illegal immigration cannot be known with certainty, it has clearly increased significantly over the past 30 years. By some estimates, the inflow of illegal immigrants has increased by a factor of five since the 1980s.

What can explain this long-term trend of increasing illegal immigration? Large wage differentials encourage illegal immigration (given limits on legal immigration) while enforcement of immigration law discourages it. These factors alone, however, cannot explain its long-term rise. One explanation is that migration is encouraged by the spread of social networks. When migrants from Mexico arrive in the United States, they are able to find friends and family from Mexico who welcome them, help them find jobs and housing, and otherwise facilitate their adaptation to the United States. Illegal immigration is thereby self-sustaining. It tends to grow over time because it spreads and deepens the social networks that facilitate it. (It is also worth noting that this growth of illegal immigration explains the increasing intensity of the controversy over it, a point we return to below.)

## Macroeconomic Theory

### Gravity Models

Gravity models, used in migration and trade flow literature, are used by economists to assess and predict aggregate migrant flows between pairs of locations. This is in contrast to the individual migrant decision-making models above. Gravity models borrow techniques from physics, and the applicability of a gravity model of migration depends on the relevance of the assumption that migrations of people follow laws similar to gravitational pulls. In gravity models, migration is assumed to move inversely with distance and positively with the size of an economy (squared), often measured by population size.

Modified gravity models characterize the recent literature following this technique. For example, Karemera, Oguledo, and Davis (2000) add sending and receiving country immigration regulations to the traditional gravity framework, and Lewer and Van den Berg (2008) include a measure of relative destination and sending country per capita income and show how the effects of supplementary variables on immigration to the traditional gravity model can be estimated in an augmented framework.

## Applications and Empirical Evidence

Econometric discrete choice modeling, a type of multiple regression, is a common technique used by economists to quantify determinants of migration for various populations of study, and econometric selection models have been used to study the effects of migration on economic outcomes. Some of these applications will be discussed here.

### Intranational Migration

Literature on the determinants of intranational migration suggests that life cycle considerations (e.g., age, education, family structure) and distance are key predictors of internal migrant flows (Greenwood, 1997). A complication to the unitary model of migration is that family units often migrate together, and therefore migration decision making may occur at the family level as opposed to the individual. Specifically, families may maximize family welfare as opposed to individual welfare with some family members suffering losses as the result of migration and others realizing offsetting gains.

In addition, locational attributes and amenities (disamenities), both environmental and those that are the result of public policy and market conditions, have been shown to attract (repel) internal migrants. Significant effort in economics literature has been made to examine the possibility of welfare migration or migration in response to differences in public aid availability and generosity across locations. McKinnish (2005, 2007), for example, in her examination of internal migration between U.S. counties, finds that having a county neighbor with lower welfare benefits increases welfare expenditures in border counties relative to interior counties. Welfare migration may occur among both those native to a country and new to it.

### International Migration

#### *The Immigration Debate in the United States*

Intranational migration is rarely controversial. In contrast, international migration often arouses heated controversies and inflammatory rhetoric. That may seem odd in the context of the United States since we like to think of ourselves as a "nation of immigrants." Why are there such passionate arguments about people who seem to like our country so much that they want to move here?

It is worth taking a moment to address this question since it puts the more technical issues in a larger, interpretive framework. The most direct answer concerns the sheer number of people who come to the United States each year. Since 2000, an average of about one million legal immigrants (Department of Homeland Security, 2008) and about 700,000 illegal immigrants (Passel & Cohn, 2008a) have entered the United States each year. About 300,000 foreigners have left the United States each year (Shrestha, 2006). Thus, net immigration has been directly increasing the U.S. population by 1.4 million persons per year. Net immigration then has indirect, subsequent effects on population growth due to immigrant fertility.

Taken together, the direct and indirect impacts of immigration on the U.S. population are startling. The Census Bureau projects that the total U.S. population will grow to 439 million by 2050, an increase of 157 million, or 56%, since 2000. To put this in concrete terms, this is equivalent to adding the entire populations of Mexico and Canada to today's (2009) population of the United States. According to Passel and Cohn (2008b), over four fifths of that growth will be due to immigrants and their descendents. Thus, immigration is dramatically increasing the number of people living in the United States.

Rapid population growth puts stress on society, on the environment, on the economy, on schools and neighborhoods, and on government. More people means more pollution, more crime, more crowding, and more need for government services. Americans take pride in their immigrant history, but they are also concerned about the impact of large-scale immigration, particularly when much of it seems to be illegal and uncontrollable. They are empathetic with immigrants, but they also are concerned about their own citizens and their own national identity. That conflict explains the intensity of the debate about U.S. immigration policy. Americans are caught between competing ideals, and neither side of the debate is obviously right.

Of all the concerns raised by mass immigration, its economic impact is perhaps the most complicated. Immigration has both good and bad effects on the economy and the workers of the host country that can be difficult to separate out. On one hand, immigration adds to labor resources and thus to the capacity for economic growth. Growth increases national income and potentially raises living standards. Immigrants "take jobs that Americans do not want" (as it is commonly said) and produce goods and services that otherwise would not be produced; that generates income for Americans as well as for immigrants. The capacity for immigration to increase the incomes of natives will be especially strong for the employers who hire the immigrants and for the more highly educated native workers whose skills complement the immigrants (in effect, the two groups establish a division

of labor that benefits both). This is one basis for the claim that immigration is beneficial for the host country.

On the other hand, large-scale immigration can also harm native workers. It is a classic labor supply shock: It increases the competition for jobs and thus drives down employment and wages for native workers, especially those whose skills are most similar to those of the immigrants. Figure 68.1 illustrates a labor supply shock corresponding to large-scale immigration. In response to the shock on the labor supply side only, equilibrium labor increases from $L_1$ to $L_2$ while wages decrease from $w_1$ to $w_2$. Immigration, however, may have unpredictable effects on wages if labor demand also changes. Panel (a) of Figure 68.1 illustrates where a demand increase less than offsets the negative wage effect from increased labor supply. The final wage, $w_3$, is lower than the initial equilibrium wage, $w_1$, despite the demand increase. Panel (b) shows that the same model under different conditions, however, may yield the opposite.

Because immigrants (especially illegal immigrants) to the United States are likely to have low levels of education—one third of foreign-born persons in the United States and almost two thirds of persons born in Mexico lack a high school degree (Pew Hispanic Center, 2009)—they compete most directly with native workers who have not attended college and especially with those who have not completed high school. These native workers are likely to earn low wages and to be on the bottom of the income distribution. African Americans, Hispanic Americans, and immigrants are especially likely to fall into this category. Thus, immigration tends to be most harmful for low-wage native workers and ethnic minorities.

Of course, immigration can simultaneously increase both job competition and economic growth. Because the former tends to harm low-income natives and the latter tends to help high-income natives, the result is an increase in inequality. In other words, immigration can have regressive effects on the income distribution of natives. The economy may grow, but the benefits of that growth flow away from the bottom and toward the top.

Because the rise in income for some natives is offset by the decline for others, the aggregate effects can wash out. These countervailing tendencies are typical of immigration. Consider the impact on productivity: Immigration can increase productivity growth if it stimulates economic activity and investment, but it can reduce productivity if the surplus of low-wage workers discourages the substitution of capital for labor. Or consider the impact on government budgets: Immigrants decrease budget deficits because they pay taxes, but they increase budget deficits because they use services. Because immigration can have contradictory effects that can cancel each other out, it may be more revealing to focus on its impact on specific groups of natives rather than on the United States as a whole.

### *Immigration and Job Competition*

Perhaps the key issue at stake in the debate about immigration is the degree of job competition. The claim that immigrants "take jobs that Americans do not want" reflects a misunderstanding of microeconomics. The extent to which Americans want jobs (the labor-leisure trade-off) is a function of their pay. If labor supply shocks created by immigration drive down the wages in these jobs so that native workers leave them, it does not follow that Americans do not want these jobs. In the absence of immigration, wages would rise and American workers would be drawn back into them.

One way of examining this issue is to compare the occupational distribution of immigrants to natives. The

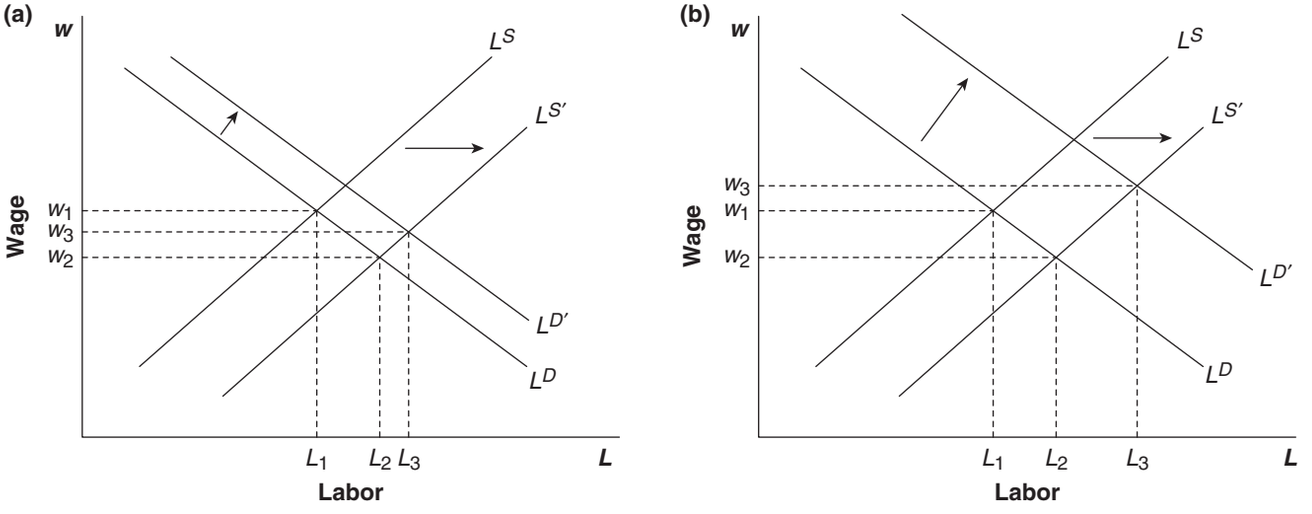


**Figure 68.1** Effects of Immigration on Equilibrium Wages

NOTES: (a) Case where labor demand increase does not offset negative wage effects from increased labor supply. (b) Case where labor demand increase more than offsets negative wage effects from increased labor supply.

two show significant overlap. According to the Pew Hispanic Center (2009), fewer than 1 in 20 foreign-born Hispanics are in agricultural or related occupations. While illegal Hispanic workers are likely undercounted in these estimates, this still suggests that the common perception that most Hispanic immigrants work in the fields—the classic "jobs that Americans do not want"—is mistaken. Most Hispanic immigrants work in occupations that native workers also hold. For example, approximately 1 in 3 foreign-born Hispanics are in construction and maintenance occupations. Their occupational distribution is more skewed toward low-wage work than non-Hispanics, but there is still considerable occupational overlap, particularly with respect to low-wage jobs. This is not consistent with the notion that immigrants and natives are in separate labor markets (a difficult notion to reconcile with the high degree of mobility and fluidity in the U.S. labor market).

To the extent that the skills and occupations of immigrant and Hispanic workers overlap, the employment and wages of native workers will be depressed by competition with immigrant workers. That is the finding of a number of researchers, most notably George Borjas (2003, 2006) and Borjas and Katz (2007). These authors find that a 10% increase in the labor supply created by immigration (about equal to the increase from 1980 to 2000) depresses the wages of all native workers by about 3% and depresses the wages of native workers without a high school degree by about 8%. These percentages are economically significant, and Borjas (2003) concludes that "immigration has substantially worsened the labor market opportunities faced by many native workers" (p. 1370).

Borjas (2003) describes his results as consistent with the simple textbook model of wage determination. However, other researchers have concluded that immigration has virtually no impact on the wages of native workers. David Card (2005) shows that the wages of native high school dropouts have not declined relative to the wages of native high school graduates since 1980, as one would expect if immigration had had the most negative impact on the wages of the least educated American workers. He concludes that "the evidence that immigrants harm native opportunities is slight" (p. F300).

Card's (2005) evidence is indirect. Instead of looking at the impact of immigration on the wages of native workers without a high school degree, he looks at their wages relative to the wages of native workers with a high school degree. The finding that their relative wages have not declined is surprising given that nationally, the real hourly wage of male and female high school dropouts fell from 1990 to 2005 while it rose for all other educational groups (Mishel, Bernstein, & Shierholz, 2009).

Card (2005) speculates that native workers are not displaced by immigrants because firms grow and invest in labor-intensive technologies in response to the availability of immigrant workers. Firms may be more likely to move to cities with large immigrant populations that make it easy to hire low-wage workers. As a result, labor demand can increase and offset the immigrant-induced increase in labor supply. Panel (b) of Figure 68.1 illustrates this case. In the figure, a demand increase more than offsets the negative wage effect from increased labor supply. The final wage, $w_3$, is higher than the initial equilibrium wage, $w_1$. Of course, depending on the magnitude of the demand shift, cases in which wages remain below $w_1$ (but are greater than $w_2$) also are possible.

This argument illustrates the capacity of the economic system to mitigate the impact of shocks over time. Immigration drives down wages over the short run by boosting labor supply, but that effect dissipates over the long run due to offsetting increases in labor demand. Other long-run adjustments demonstrate the same tendency. For example, native workers who face increasing job competition from immigrants may respond by moving to places where there are fewer immigrants. That will reduce labor supply and thus the measured impact of immigration on wages in the cities that the native workers have left (by 40% to 60% according to Borjas, 2006). Also, the native workers most marginally connected to the labor force may drop out of the labor force in response to competition from immigrants (Johannsson, Weiler, & Shulman, 2003). That takes the lowest wage native workers out of the sample, thereby raising the average wage of the workers who remain (this is similar to the argument made by Butler & Heckman [1977] about the impact of welfare on average black labor force participation and wages).

It is important to note that the capacity of the economic system to adjust to shocks over the long run does not negate the losses that have occurred and accumulated over a succession of short-run states. The short-run losses continue to occur as immigrant inflows continue. These inflows tend to rise and fall with the state of the economy and the degree of enforcement of immigration law; nonetheless, they are projected to continue in large part indefinitely. If that projection proves to be true, the "short-run" losses from immigration-induced labor supply shocks will never cease even if each shock tends to diminish over time. In this sense, both Card and Borjas can be right.

### *Immigration, Growth, and Inequality*

In the United States, immigration lowers the wages and employment of low-wage workers, but it also increases economic growth and national income. The recipients of this additional income—aside from the immigrants themselves—are the employers of the immigrants and the workers whose skills are complementary to the immigrants or who provide services to the immigrants. Since these tend to be high-skill, high-wage workers and since large employers typically have high incomes, it follows that many beneficiaries of immigration to the United States are relatively affluent.

Much of the above analysis refers to the immigration of low-skill workers. Of course, many other immigrants are highly skilled. Immigrants are almost as likely as

natives to hold a bachelor's degree and are somewhat more likely to hold a graduate degree (Pew Hispanic Center, 2009). Highly skilled immigrants bring technical and entrepreneurial skills into the United States that add to productivity, innovation, and growth. One study shows that a 1% increase in the share of immigrant college graduates in the population raises patents, relative to population, by 6% (Hunt & Gauthier-Loiselle, 2008). However, high-skill immigrants can create job competition for high-skill natives, just as low-skill immigrants can create job competition for low-skill natives. Borjas (2005) estimates that a 10% immigration-induced increase in the supply of doctorates lowers the wages of comparable natives by 3%.

Immigration also affects prices, and that also should be factored into the discussion of its overall economic consequences. Although the low wages earned by the majority of immigrants can harm the workers who compete with them, this lowers the prices of the goods and services they produce. This unambiguously benefits native consumers. Cortes (2008), for example, finds that a 10% increase in the share of low-skilled immigrants in the labor force decreases the price of immigrant-intensive services (e.g., housekeeping and gardening), primarily demanded by high-income natives, by 2%. This is suggestive of surplus gains resulting from depressed prices as an additional result of immigration.

Immigration thus has a variety of both positive and negative effects on U.S. natives. Since the negative effects tend to be concentrated among low-income natives, and since the positive effects may be concentrated among high-income natives, there are two overall consequences. First, immigration may increase inequality along with other trends such as technological change, the decline in unions, international competition, and deindustrialization. The extent (and existence) of that increase in inequality is a matter of dispute, as one would expect from the corresponding dispute about the impact of immigration on the wages of low-skill native workers. Borjas, Freeman, and Katz (1997) conclude in a widely cited paper that immigration accounts for a quarter to a half of the decline in the relative wages of low-skill native workers. That would make it a significant factor in the increase in inequality between low-wage and high-wage workers. In contrast, a more recent paper by Card (2009) concludes that immigration does not significantly increase wage inequality.

Second, the gains and losses from immigration may approximately cancel out in the aggregate. This can be understood in two ways. First, the gains that accrue to affluent Americans are offset by the losses to low-income Americans; second, the benefits created by high-skill immigrants are offset by the losses created by low-skill immigrants. Consequently, the estimates of the aggregate impact of immigration are small relative to the size of the U.S. economy. The study of immigration conducted by the National Research Council (1997) concluded that immigration adds at most about 0.1% to gross domestic product (GDP).

Even that small benefit gets wiped out by the net fiscal costs imposed by immigration. Immigrants are disproportionately likely to have low incomes and large families. Thus, they have relatively high needs for social services (particularly schools) but pay relatively little in taxes. The fiscal balance is positive at the federal level because the federal government receives the Social Security taxes of immigrants but provides little in the way of services or Social Security benefits; however, it is negative at the state and local levels, substantially so in the locales most heavily affected by immigration. The negative effect outweighs the positive effect by a small amount. The study just cited showed that the current fiscal burden imposed by immigration is only about $20 per household as of the mid-1990s, but added up over all households, it amounts to an overall loss of about the same amount as immigration adds to economic growth. It therefore seems safe to conclude that the overall economic effect of immigration is approximately zero. In this sense, both sides of the immigration debate (one claiming that immigration is an economic disaster and the other claiming that immigration is an economic necessity) are wrong.

These conflicting considerations do not lend themselves to simple conclusions about the impact of immigration. Overall judgments will depend upon which group we view with most concern. For example, immigrants clearly benefit from immigration, so those who care most about the well-being of immigrants will support a more expansionist approach to immigration. Those who care most about low-income natives would support a more restrictionist approach. Those who care most about the business sector would support a more expansionist approach. Other policy combinations could arise as well. For example, those who care about both immigrants and low-income natives might support a more expansionist approach combined with government assistance to the adversely affected natives. Or those who care most about both immigrants and low-income natives might support a more restrictionist approach combined with government foreign aid to raise standards of living and reduce the incentive to emigrate from sending countries. These complicated ethical and political judgments cannot be resolved just by recourse to the economic evidence.

### *Immigrant Assimilation*

In addition to the economic effects on U.S.-born workers, another area of debate concerns economic assimilation or whether immigrants' earnings distributions approach those of natives as time elapses within a host country. Recent empirical evidence generally does not support full economic assimilation by immigrants. Cross-sectional regressions in the early literature predicted rapid increases in immigrant earnings upon arrival in the United States. Borjas (1985), however, found that within-cohort growth is significantly smaller than previous estimates

using cross-sectional techniques. In a follow-up, Borjas (1995) found evidence that increases in relative wages of immigrants after arrival in the United States are not enough to result in wages equivalent to those of like natives. Instead, immigrants earn 15% to 20% less than natives throughout most of their working lives.

There also are mixed results regarding assimilation across generations. Some authors argue that intergenerational assimilation may be faster than assimilation of migrants themselves. Card (2005), for example, argues that while immigrants themselves may not economically assimilate completely, children of immigrants will join a common earnings path with the children of natives, and thus assimilation is an intergenerational process. Therefore, intergenerational assimilation may be faster than assimilation of immigrants themselves. Smith (2003) finds that each successive generation of Hispanic men has been able to decrease the schooling gap, and this has translated into increased incomes for subsequent generations. On the other hand, some authors have argued that Mexican immigrants have slower rates of assimilation in comparison to other immigrants. Lazear (2007), for example, argues that immigrants from Mexico have performed worse and become assimilated more slowly than immigrants from other countries. He argues that this is the result of U.S. immigration policy, the large numbers of Mexican immigrants relative to other groups, and the presence of ethnic enclaves.

### *Immigrant Locational Choice*

A number of state economic and demographic conditions and state policy instruments can be hypothesized to affect the locational distribution of immigrants. Bartel (1989) studies the locational choices of post-1964 U.S. immigrants at the city level and finds that immigrants are more geographically concentrated than natives while controlling for age and ethnicity and that education reduces the probability of geographic clustering and increases the probability of changing locations after arrival in the United States. Jaeger (2000) finds that immigrants' responsiveness to labor market and demographic conditions differs across official U.S. immigrant admission categories, including admission based on presence of U.S. relatives, refugee or asylee status, and employment or skills. Employment category immigrants, for example, are more likely to locate in areas with low unemployment rates. Other determinants of locational choice are wage levels and ethnic concentrations. A growing literature has examined how information networks affect migration decisions and outcomes conditional on arrival. Munshi (2003), for example, finds that Mexican immigrants with larger networks are more likely to be employed and to hold a higher paying nonagricultural job.

As in studies of internal migration, the possibility of welfare migration has been another theme in economic literature on the effects of international migration. If immigrants choose locations based on welfare availability and generosity, there may be fiscal consequences to certain communities. Like the literature on the effects of immigration overall, the literature on immigrant welfare migration is characterized by debates over appropriate data sources, econometric methods, and ultimate results.

Borjas (1999) presents a model of whether welfare-generous states induce those immigrants at the margin (who may have stayed home or located elsewhere in the absence of welfare) to make locational decisions based on social safety net availability. He examines whether the interstate dispersion of public service benefits influences the locational distribution of legal immigrants relative to the distribution of U.S.-born citizens. In this framework, he demonstrates that immigrant program participation rates are more sensitive to benefit-level changes than native participation rates are. Conclusions here, however, are sensitive to specification and the particular data source used. While some other authors find strongly positive, significant relationships between legal immigration flows and welfare payment levels (e.g., Buckley, 1996; Dodson, 2001), others conclude the opposite (e.g., Kaushal, 2005; Zavodny, 1999). Whether immigrants locate based on welfare generosity and availability is still a subject of academic debate within economics.

### *Effects on Sending Countries: Brain Drain, Remittances, and Intergenerational Effects*

The effects of immigration on sending and receiving countries depend on how income distributions compare and whether immigrants come from the high or low end of the skill distribution within the sending economy (Borjas, 1987). If immigrants come from the high end of the skill distribution, then we may characterize this phenomenon as positive selection. Likewise, if immigrants come from the low end of the skill distribution, then we would refer to this as negative selection. Positive or negative selection may result from differences in the rate of return to skill across locations. If immigrants are positively selected from the sending country, then economists would describe the country as experiencing brain drain.

Recent studies have suggested positive, not negative, effects on sending countries overall. Beine, Docquier, and Rapoport (2001), for example, describe two effects of migration on human capital formation and growth in a small, open developing economy. First, migration leads to increases in the demand for education as potential migrants predict higher returns abroad, and second, brain drain occurs as people actually migrate. The net effect on the sending country depends on the relative magnitudes of these effects, and cross-sectional data for 37 developing countries suggest that the first effect may dominate the second. Stark (2004) presents similar results from a model where a positive probability of migration is welfare enhancing in that it raises the level of human capital in the sending country.

In addition to changing the skill distribution in host and sending countries, migrants may affect the cross-country income distribution by remitting portions of their income abroad to family members in the sending country. These remittances may decrease the prevalence of poverty in some locations in the sending country and therefore represent additional welfare improvements. The quantity of remittances, however, may be interrelated with the presence of brain drain. Faini (2007), for example, finds evidence that brain drain is associated with a smaller propensity to remit. Although skilled migrants may make higher wages in the second country, they are empirically found to remit less than do unskilled migrants.

Other areas of literature relating to the economics of immigration examine intergenerational effects of migration. Of particular interest, increasing research has focused on the effects of immigration on subsequent generations both in host and sending countries.

*Illegal Immigration and Repeat Migration*

As may be expected, research on undocumented or illegal immigration is plagued by a lack of reliable and representative data. Estimates on the aggregate size of the illegal foreign-born population generally rely on residual methodology. Passel (2006a, 2006b) estimates that 11.1 million illegal immigrants were present in the United States in March 2005. This estimate is up from 9.3 million in 2002 (Passel, 2004) and 10.3 million in 2003 (Passel, Capps, & Fix, 2005). Of this total, approximately 6.2 million (56%) were from Mexico. In terms of spatial distribution within the country, 2.5 to 2.75 million illegal immigrants resided in California, followed by Texas (1.4–1.6 million), Florida (0.8–0.95 million), New York (0.55–0.65 million), and Arizona (0.4–0.45 million). These five states account for more than 50% of the estimated total. Furthermore, some immigration streams are characterized by repeat migration where, for example, migrants work for a short period (sometimes a season) before returning to their country of origin and repeat this cycle several times. Seasonal agricultural work in the United States, for example, takes this pattern where migrants (particularly from Mexico) work a season, return to their country of origin, and then return the following year.

Evidence on whether border enforcement affects the locational distribution of illegal immigrants is mixed. Some authors conclude that border enforcement causes migrants to make several attempts to cross the border as opposed to deterring migration, that the composition of illegal migrants may respond to increases in border patrol, and that the distribution of destinations may be sensitive to border patrol intensity. Others do find a deterrence effect. Orrenius (2004), for example, considers preferred border crossing sites at the state and city levels and concludes that enforcement has played an important role in deterring Mexican migrants from crossing in California. Gathmann (2008) and Dávila, Pagán, and Soydemir (2002) present results supporting this conclusion. Hanson (2006) summarizes existing literature on illegal immigration from Mexico to the United States and points to areas for future research, particularly that pertaining to policies to control labor flows.

## Policy Implications

### History of U.S. Immigration Policy and Current Reform Proposals

Recent U.S. immigration debates have included proposals for legalization or amnesty of long-term undocumented immigrants. The Comprehensive Immigration Reform Act of 2006 (CIRA), for example, suggested the admission of undocumented immigrants present in the United States more than 5 years, subject to the condition that these persons pay fines and back taxes. Those with 2 to 5 years of U.S. tenure would be allowed to remain in the country for 3 years, after which they would return to their countries of origin to apply for citizenship. In addition, CIRA proposed new guest worker programs. The 1986 Immigration Reform and Control Act (IRCA) was similar. IRCA included a general program (I-687) granting legal status on the basis of continuous U.S. residence for the 5 years leading up to the program and the Seasonal Agricultural Worker (SAW) program (I-700) for farm workers employed at least 90 days in the previous year. Applications for amnesty totaled 1.8 million under I-687 and 1.3 million under I-700, and a total of 2.7 million received legal permanent residency. Applicants were first granted temporary resident status, followed by permanent residency after passing English-language and U.S. civics requirements.

Understanding the effects of legal status on worker outcomes is crucial to anticipating potential effects of a new amnesty program. Several studies document a wage gap between documented and undocumented immigrants. Borjas and Tienda (1993), using administrative data for IRCA amnesty applicants, construct wage profiles by legal status and find that documented immigrants earn 30% more than undocumented immigrants with like national origins. Rivera-Batiz (1999), using a short panel of IRCA amnesty recipients, finds that average hourly wage rates of Mexican documented immigrants were approximately 40% higher than those of undocumented workers at the time amnesty was granted. Decomposing this wage differential into explained and unexplained parts, he finds that less than 50% of this gap is attributable to differences in observed worker characteristics. Furthermore, he confirms that undocumented immigrants who received amnesty in 1987 or 1988 had economically and statistically significant increases in earnings of the order of 15% to 21% by the follow-up survey in 1992. Less than 50% of this increase can be explained

by changes in measured worker characteristics over this time period. Kossoudji and Cobb-Clark (2002) find the wage penalty for being undocumented to be in the range of 14% to 24% and estimate that the wage benefit that accrued to those legalized under IRCA was around 6%. Pena (2010) estimates wage differentials between legal and illegal U.S. farm workers to be on the order of 5% to 6%. Whether workers would realize this full gain as a result of legalization, however, depends on a number of general equilibrium characteristics and is unlikely.

Amnesties have been controversial because they seem to reward lawbreaking and because they can create incentives for more illegal immigration. Although CIRA passed the U.S. Senate in May 2006, it failed to pass the House of Representatives. Alternatives to amnesty include continued increases in border security via border enforcement efforts and extensions to temporary work permit programs. These alternative proposals suggest areas for future research regarding predicted effects of immigration reform on U.S. labor markets.

## Conclusion

The economics of migration is an application of microeconomic and macroeconomic theory. Still, given the complexity of interrelationships between host and sending countries, many questions of migration dynamics and of the causes and effects of internal and international migration are empirical ones and are the subject of debates over data sources, methodology, and conclusions. Empirical arguments such as these are hardly unusual in the social sciences. Far from being a reason for cynicism, scholarly debates are progressive—some issues get resolved while others emerge and continue to be debated—and are necessary for rational policy making. Yet it is important to note their limitations as well. The economic analysis of migration cannot answer some of the basic questions at the heart of the public debates about immigration, such as which and how many people should be allowed into a country, how their interests should be balanced against the interests of natives, and how and to what extent national identity should be preserved. These ethical and political issues can be informed by economic analysis, but ultimately they cannot be resolved by it.

## References and Further Readings


Bartel, A. P. (1989). Where do the new U.S. immigrants live? *Journal of Labor Economics, 7,* 371–391.

Beine, M., Docquier, F., & Rapoport, H. (2001). Brain drain and economic growth: Theory and evidence. *Journal of Development Economics, 64,* 275–289.

Borjas, G. J. (1985). Assimilation, changes in cohort quality, and the earnings of immigrants. *Journal of Labor Economics, 3,* 463–489.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *American Economic Review, 77,* 531–553.

Borjas, G. J. (1995). Assimilation and changes in cohort quality revisited: What happened to immigrant earnings in the 1980s? *Journal of Labor Economics, 13,* 201–245.

Borjas, G. J. (1999). Immigration and welfare magnets. *Journal of Labor Economics, 17,* 607–637.

Borjas, G. J. (2003). The labor demand curve is downward sloping: Examining the impact of immigration on the labor market. *Quarterly Journal of Economics, 118,* 1335–1374.

Borjas, G. J. (2005). The labor-market impact of high-skill immigration. *American Economic Review, 95*(2), 56–60.

Borjas, G. J. (2006). Native internal migration and the labor market impact of immigration. *Journal of Human Resources, 41,* 221–258.

Borjas, G. J., Freeman, R. B., & Katz, L. F. (1997). How much do immigration and trade affect labor market outcomes? *Brookings Papers on Economic Activity, 1,* 1–90.

Borjas, G. J., & Katz, L. (2007). The evolution of the Mexican-born workforce in the United States. In G. J. Borjas (Ed.), *Mexican immigration to the United States* (pp. 13–55). Chicago: University of Chicago Press.

Borjas, G. J., & Tienda, M. (1993). The employment and wages of legalized immigrants. *International Migration Review, 27,* 712–747.

Buckley, F. (1996). The political economy of immigration policies. *International Review of Law and Economics, 16,* 81–99.

Butler, R., & Heckman, J. (1977). The government's impact on the labor market status of black Americans: A critical review. In L. Hausman, O. Ashenfelter, B. Rustin, R. F. Schubert, & D. Slaiman (Eds.), *Equal rights and industrial relations* (pp. 235–281). Madison, WI: Industrial Relations Research Association.

Card, D. (2005). Is the new immigration really so bad? *The Economic Journal, 115*(507), F300–F323.

Card, D. (2009). *Immigration and inequality* (NBER Working Paper No. 14683). Cambridge, MA: National Bureau of Economic Research.

Cortes, P. (2008). The effect of low-skilled immigration on U.S. prices: Evidence from CPI data. *Journal of Political Economy, 116,* 381–422.

Dávila, A., Pagán, J. A., & Soydemir, G. (2002). The short-term and long-term deterrence effects of INS border and interior enforcement on undocumented immigration. *Journal of Economic Behavior and Organization, 49,* 459–472.

Department of Homeland Security, Office of Immigration Statistics. (2008). *2007 yearbook of immigration statistics.* Washington, DC: Author.

Dodson, M. E., III. (2001). Welfare generosity and location choices among new United States immigrants. *International Review of Law and Economics, 21,* 47–67.

Faini, R. (2007). Remittances and the brain drain: Do more skilled migrants remit more? *The World Bank Economic Review, 21,* 177–191.

Gathmann, C. (2008). Effects of enforcement on illegal markets: Evidence from migrant smugglers along the southwestern border. *Journal of Public Economics, 92,* 10–11, 1926–1941.

Greenwood, M. J. (1997). Internal migration in developed economies. In M. R. Rosenzweig & O. Stark (Eds.), *Handbook of population and family economics* (pp. 647–720). Amsterdam: North-Holland.

Hanson, G. H. (2006). Illegal migration from Mexico to the United States. *Journal of Economic Literature, 44,* 869–924.

Hunt, J., & Gauthier-Loiselle, M. (2008). *How much does immigration boost innovation?* (NBER Working Paper No. W14312). Cambridge, MA: National Bureau of Economic Research.

Jaeger, D. A. (2000). *Local labor markets, admission categories, and immigrant location choice.* Unpublished manuscript, College of William & Mary, Williamsburg, VA.

Johannsson, H., Weiler, S., & Shulman, S. (2003). Immigration and the labor force participation of low-skill workers. In S. Polachek (Ed.), *Worker well-being and public policy: Research in labor economics* (Vol. 22, pp. 291–308). New York: JAI.

Karemera, D., Oguledo, V. I., & Davis, B. (2000). A gravity model analysis of international migration to North America. *Applied Economics, 32,* 1745–1755.

Kaushal, N. (2005). New immigrants' location choices: Magnets without welfare. *Journal of Labor Economics, 23,* 59–80.

Kossoudji, S. A., & Cobb-Clark, D. A. (2002). Coming out of the shadows: Learning about legal status and wages from the legalized population. *Journal of Labor Economics, 20,* 598–628.

Lazear, E. P. (2007). Mexican assimilation in the United States. In G. J. Borjas (Ed.), *Mexican immigration to the United States* (pp. 107–122). Cambridge, MA: National Bureau of Economic Research.

Lewer, J. J., & Van den Berg, H. (2008). A gravity model of immigration. *Economics Letters, 99,* 164–167.

McKinnish, T. (2005). Importing the poor: Welfare magnetism and cross-border welfare migration. *Journal of Human Resources, 40,* 57–76.

McKinnish, T. (2007). Welfare-induced migration at state borders: New evidence from micro-data. *Journal of Public Economics, 91,* 437–450.

Mishel, L., Bernstein, J., & Shierholz, H. (2009). *The state of working America 2008–2009.* Ithaca, NY: Cornell University Press.

Munshi, K. (2003). Networks in the modern economy: Mexican migrants in the U.S. labor market. *Quarterly Journal of Economics, 118,* 549–599.

National Research Council. (1997). *The new Americans: Economic, fiscal and demographic effects of immigration.* Washington, DC: National Academy Press.

Orrenius, P. M. (2004). The effect of U.S. border enforcement on the crossing behavior of Mexican migrants. In J. Durand & D. S. Massey (Eds.), *Crossing the border: Research from the Mexican Migration Project* (pp. 281–298). New York: Russell Sage Foundation.

Passel, J. S. (2004). *Undocumented immigrants: Facts and figures.* Washington, DC: Urban Institute, Immigration Studies Program.

Passel, J. S. (2005). *Unauthorized migrants: Numbers and characteristics.* Washington, DC: Pew Hispanic Center.

Passel, J. S. (2006a). *Estimates of the unauthorized migrant population for states based on the March 2005 CPS.* Washington, DC: Pew Hispanic Center.

Passel, J. S. (2006b). *The size and characteristics of the unauthorized migrant population in the U.S.: Estimates based on the March 2005 Current Population Survey.* Washington, DC: Pew Hispanic Center.

Passel, J. S., Capps, R., & Fix, M. (2005). *Estimates of the size and characteristics of the undocumented population.* Washington, DC: Pew Hispanic Center.

Passel, J. S., & Cohn, D. (2008a). *Trends in unauthorized immigration: Undocumented inflow now trails legal inflow.* Washington, DC: Pew Hispanic Center.

Passel, J. S., & Cohn, D. (2008b). *U.S. population projections: 2005–2050.* Washington, DC: Pew Hispanic Center.

Pena, A. A. (2010). Legalization and immigrants in U.S. agriculture. *The B. E. Journal of Economic Analysis & Policy, 10*(1), Article 7.

Pew Hispanic Center. (2009). *Statistical portrait of the foreign-born population in the United States, 2006.* Washington, DC: Author.

Rivera-Batiz, F. L. (1999). Undocumented workers in the labor market: An analysis of the earnings of legal and illegal Mexican immigrants in the United States. *Journal of Population Economics, 12,* 91–116.

Shrestha, L. (2006). *The changing demographic profile of the United States* (CRS Report RL 32701). Washington, DC: Congressional Research Service.

Sjaastad, L. A. (1962). The costs and returns of human migration. *Journal of Political Economy, 70*(5), 80–93.

Smith, J. P. (2003). Assimilation across the Latino generations. *American Economic Review, 93,* 315–319.

Stark, O. (2004). Rethinking the brain drain. *World Development, 32,* 15–22.

Tiebout, C. M. (1956). A pure theory of local expenditures. *Journal of Political Economy, 64,* 416–424.

Zavodny, M. (1999). Determinants of recent immigrants' locational choices. *International Migration Review, 33,* 1014–1030.

# 69

# Health Economics

Shirley Johnson-Lans
*Vassar College*

Health economics is widely understood to encompass the study of the demand and supply for medical services (physician services, services provided in hospitals and independent laboratories, pharmaceuticals, etc.) and for health insurance, as well as comparative studies of different health care systems. It also includes the study of the determinants of demand for health itself, global public health problems, and the nonmedical inputs into health, such as a decent living standard, education, physical and social environment, and personal lifestyle choices, to the extent that they are exogenous (e.g., independent of one's health status). Although the nonmedical factors are increasingly realized to be important in achieving a healthy community at an affordable level of expenditure, most courses in health economics are primarily concerned with the provision of medical care and with health insurance that primarily covers medical care. This chapter will adhere to that tradition since expenditure on medical care, insurance, and research represents such a high proportion of gross domestic product (GDP), especially in the United States, and a proportion that is increasing in all high-income industrialized nations. We also focus on medical care as an input into health because it provides no benefits other than its contribution to health, unlike diet, recreation, and exercise.

Medical care also differs from most other expenditures, even those that we think of as human capital investments, because much of it occurs as a result of negative shocks to health that are largely unanticipated. It is the combination of the degree of uncertainty about one's future health state and the high cost of medical care (relative to household budgets) that makes the transferring of risk to a third-party payer through insurance such an important phenomenon in the market for health care. For this reason, the role of health insurance will be discussed before the analysis of markets for other health care services.

The last several sections of this chapter are devoted to policy concerns. Before considering possible reforms of the U.S. health care system, a brief overview of several other countries' health care systems will be provided.

## Methodology Used in Health Economics Research

The methodology of health economics research includes the following two categories.

### Statistical Techniques

In health economics, experimental laboratory conditions rarely can be created. Therefore, once a hypothesis has been formulated and sample data have been gathered, statistical techniques must be used to isolate and estimate the effects of particular factors. Economists most commonly "remove" the other effects by using the techniques of multivariate correlation and regression analysis. Whenever possible, researchers use "difference-in-difference" estimators, where changes in a control group are compared with changes in a treatment group.

In isolating the effect of a change in policy or environment, one needs to have a control group to compare with a treatment group. In some cases, "natural experiments" are provided by the environment. For example, quasi-experimental conditions were provided when Tennessee raised its rate of Medicaid remuneration for physician visits while a neighboring state,

Maryland, did not. This enabled researchers to estimate the effect of fees on willingness of physicians to treat Medicaid patients. Physicians in Maryland were the control group.

Researchers occasionally are able to undertake experiments in which large numbers of subjects are randomly assigned to different groups. An example is the RAND Health Insurance Experiment, conducted over 1974 to 1982. More than 2,000 households were randomly assigned to a variety of insurance plans that offered differing degrees of coverage. This freed the research from the problem of selection bias that results when people systematically choose different insurance plans based on their expected use of medical care. Today an opportunity for a new experiment has been provided by a decision of Medicaid in Oregon to establish a lottery to randomly choose people from the eligible pool who will receive coverage for medical care.

### Cost-Benefit and Cost-Effectiveness Analysis

Cost-benefit analysis is a strategy for comparing benefits with costs. It compares marginal benefits and marginal costs and employs the rule that one should devote resources to a use until the extra or incremental cost of the last unit just equals the incremental benefit of that unit. *This rule assumes that marginal benefits are declining and marginal costs are either constant or rising.* The underlying assumption is that people are rational and thus want to maximize benefits relative to costs. This approach is used to answer whether an activity is worth undertaking or continuing. It can be used only if we can measure both costs and benefits in the same metric. We can only decide whether the cost of a medical treatment is worth it if we can establish a monetary value for the benefit. For example, when considering whether to undergo heart surgery, the probable effects of the heart surgery are usually stated in terms of an estimated improvement in length of life and/or quality of remaining life years. To use cost-benefit analysis, one must assign a monetary value to a year of life or to a given quality-of-life improvement.

When it is not possible to establish monetary values for benefits, cost-effectiveness analysis may be used to compare marginal costs, expressed in monetary terms, with incremental benefits, expressed in natural units, such as amount of improvement in life expectancy or degree of reduction in blood pressure. Cost-effectiveness analysis can never establish whether some course of action is worthwhile, but it can be used to compare different treatment methods in terms of their relative effectiveness. This analysis can only provide unambiguous results when one alternative provides at least as good an outcome using fewer resources or a better outcome using the same level of resources. It is a better indicator of a rational use of resources than just cost, whether we are considering decision making of individuals or of societies.

## The Demand for Health

It is important to clearly distinguish between health and health care. Health can be considered a form of human capital (like education), and medical care and other components of health care are inputs into the production of health. Spending on health is more appropriately treated as an investment in a stock of health (capital) rather than an item of current consumption. The formal model of investment in health, developed by Michael Grossman, employs a marginal efficiency of health capital function (MEC), which we can think of as a quasi-demand function for health (Grossman, 1972). The MEC is specific to an individual in that different people are endowed with different initial stocks of health and also suffer different shocks to their health status over time. It is downward sloping because it assumes diminishing returns to marginal inputs into the production function. Grossman distinguished between gross and net investment in health since there is depreciation in health capital that must be overcome as well as net investment in improvements in the health stock.

Things that augment the value of healthy days will increase the demand for health. Thus, an increase in the wage rate will shift the demand function. Education has also been found to be positively associated with the demand for health, although just why is still under investigation. Education may shift the demand for health because of a taste or preference change, and/or it may increase the productivity of the inputs into health.

This model can easily be accommodated to incorporate risk or uncertainty. The role of uncertainty in the demand for health and health care is very important. Uncertainty about one's future state of health, uncertainty about what kind of treatment to pursue, and uncertainty about the cost of treatment all contribute to the importance of insurance, or "third-party payment" in the market for health care services. Uncertainty has implications not only for the role of insurance but also for the relationship between patients and their health care providers, particularly their physicians (Arrow, 1963). Patients consult physicians in large part because of the latter's expertise. Asymmetry of information thus introduces the classic principal/agent set of problems that will be discussed later.

## Health Insurance

Insurance is a mechanism for assigning risk to a third party. It is also a mechanism for pooling risk over large groups, which in the case of health insurance involves transferring benefits from healthy to ill individuals within a pool. Health insurance may be either private insurance, purchased by individuals or groups, or social insurance, provided by governments out of tax revenues.

## The Demand for Private Health Insurance

Why is health insurance so widely purchased? The demand exists because people desire to protect themselves against potential financial losses associated with the treatment of illness. Why don't they self-insure themselves by saving money when they are well to use in times of illness? There are a number of reasons, including the fact that many people could never save or borrow enough to pay for potential catastrophic levels of medical expenditure. However, even people who have extensive wealth usually buy insurance. The reason is that most people want to avoid risk—that is, they are "risk averse." Economists define risk aversion as a characteristic of people's utility functions. Attitudes toward risk depend on the marginal utility of an extra dollar (lost or gained). If the marginal utility of an extra dollar is decreasing as wealth increases, a small probability of a large reduction in wealth entails a larger loss of utility than the certain loss of a smaller amount of wealth (the cost of the insurance) when the probability weighted or "expected value" of the two alternatives is equal. This is what is meant by being "risk averse." Risk-averse people will be willing to pay for insurance even though the cost of the insurance premium is more than the expected value of loss due to illness. (Insurance companies are willing to supply insurance when this excess over expected payouts allows them to cover administrative costs, build up a reserve fund, and, in the case of for-profit insurance firms, make a profit.)

## Structure of Private Health Insurance Contracts

Insurance policies are commonly structured as indemnity contracts, where individuals are compensated by a certain amount in the event of an adverse event. Historically, most health insurance policies covering hospitalization and other health care services were modified indemnity contracts in which a certain portion of fee-for-service costs of covered services was reimbursed.

Deductibles and co-payments are used by insurance companies to try to limit the degree of moral hazard associated with insurance coverage. *Moral hazard* is the phenomenon of a person's behavior being affected by insurance coverage. The main way in which moral hazard operates in the health insurance market is through the tendency for the insured to use a greater quantity of medical care since insurance lowers its cost to the individual. Breadth of coverage increases moral hazard since more "discretionary" services are included. Moral hazard is greater where the price elasticity of demand for the service is greater.

Most private health insurance in the United States is provided on a group basis at one's place of employment. This type of health insurance has been favored by workers and firms since the 1950s in part because of the favorable tax treatment it receives in the United States. The cost of employment-based health insurance is not considered taxable income to employees but can be deducted by firms as labor expense. It is also favored because larger insurance pools usually involve lower premiums.

The problem of *adverse selection* is a feature of insurance markets in which there are multiple insurance pools. A pool suffers from adverse selection if it is composed of older, sicker, or other individuals more prone to use medical services. Insurance companies will, if legally allowed, charge higher premiums to higher risk individuals, families, or groups or avoid insuring them altogether. In some cases, regulation requires insurers to use community rating (e.g., charge the same premium to all subscribers for the same coverage), regardless of risk factors such as age or medical histories. Group health insurance is one means of dealing with adverse selection since risk factors are pooled for the group and community rates are charged to all members of the group.

In the past 20 years, health insurance policies have also evolved to incorporate aspects of "managed care," a variety of features instituted by third-party payers to contain costs and provide greater efficiency in the provision of services.

An important innovation in managed care is the sharing of risk not only with consumers of health care (through deductibles, co-payments, lifetime payout limits, etc.) but also with providers of health care (by entering into contracts with them that set the amount of reimbursement per treatment or by paying a fixed amount per subscriber).

In some cases, individuals join health maintenance organizations (HMOs), which provide integrated health care services and require subscribers to use in-network providers. In HMOs, primary care physicians act as "gatekeepers"; subscribers must get referrals from them to be reimbursed for other specialized services. The most common form of managed care contract in the United States today is the preferred provider organization (PPO). In this arrangement, consumers face lower prices if they use "in-network" providers but also receive some reimbursement for other services, although these usually have higher co-payments and are subject to deductibles. The PPO is a hybrid between the traditional indemnity contract and an HMO. In PPOs, the gatekeeper function of a primary care physician is not imposed.

In managed care contracts, there are usually requirements that patients receive precertification (e.g., obtain permission from the insurance company) before surgery, diagnostic tests, and so on, and physicians and hospitals are often subject to utilization review of treatments prescribed.

As health care costs have risen over time, insurance premiums have also risen. Employers, particularly those with smaller groups of workers, have tended to cut back on coverage, require employees to pay higher proportions of premiums, and in some cases discontinue coverage. Workers, even in jobs that provide options for group insurance, have also tended, in increasing numbers, to fail to subscribe to group plans as their required contributions have risen. Thus today, the ranks of the uninsured include many workers as well as people who are unemployed, not in the labor force, or self-employed.

### Social Insurance

Social insurance, in addition to pooling risk, usually has a redistributive function since it is financed out of taxes, so one pays an amount dependent on income but receives benefits in accordance with need. In the United States, social insurance covers only certain groups: those over 65 years of age, most of whom are covered by Medicare; a portion of low-income persons who are covered by Medicaid; children of low-income families, who may be covered by the State Child Health Insurance Program (SCHIP); Native Americans living on reservations; and veterans who can receive health care services through the Veterans Administration. Certain other people with approved disabilities may also be eligible for social insurance. Federal employees, including members of Congress, are also covered by public insurance.

The two largest programs, Medicare and Medicaid, came into being in 1965 as amendments to the Social Security Act. Medicare Part A, which covers all senior citizens who are Social Security eligible, covers part of hospital bills and is financed federally out of payroll taxes paid jointly by employees and employers. Medicaid Parts B and D are voluntary and require contributions out of Social Security retirement benefit checks but are heavily subsidized. Medicaid is jointly financed by federal and state tax revenues. States administer Medicaid and are required to finance their portion of their programs or the federal contribution is reduced.

The financial solvency of Medicare is vulnerable not only to rising health care costs but also to shifts in the age distribution of the population since it is financed on a pay-as-you-go basis, which means that contributions from current workers are used to pay benefits to current beneficiaries. The financial solvency of Medicaid, which is means tested, is vulnerable to cycles in economic activity since during recessions, tax receipts fall and eligibility roles swell. The problem is exacerbated by the fact that states are required to balance their budgets annually.

Both Medicare and Medicaid originally reimbursed physicians and hospitals on a fee-for-service basis. However, both programs have adopted some of the same managed care strategies used by private insurance companies, and in fact, Medicare pioneered a prospective payment system of hospital reimbursement based on fixed payments per diagnosis. This diagnostic-related group (DRG) method has been copied by many private insurers. Physicians' payments are also set according to a scale (RBRVS) that determines reimbursement rates for different types of physician visits, factoring in the input resources (effort) used and the costs (based on capital intensity of an office practice, number of years of training required for a specialty, etc.).

Medicare and Medicaid both contract out to some private HMOs and other managed care insurers. Medicare Parts B and D contain options for subscribers to assign their premiums to private third-party payers. The intent is to provide competition in the social insurance market. Currently, the government is subsidizing the Medicare Advantage Programs (private options), paying more per beneficiary than for traditional Medicare Part B. Some states require Medicaid beneficiaries to receive services from an HMO.

The United States is nearly unique among high-income industrialized countries in not having universal health insurance coverage. About 15% of the U.S. population currently has no health insurance. It should be noted, however, that universal coverage does not mean universal social insurance coverage. Proposals for increasing insurance coverage of Americans involve various combinations of social and private insurance.

## Markets for Physicians and Nurses, Hospitals, and Pharmaceuticals

### Supply and Demand for Physicians and Registered Nurses

*Physicians*

The supply of professionals, whether physicians, nurses, engineers, attorneys, or accountants, depends on the willingness of people to undertake training to enter these professions. This investment in human capital is determined, at least in part, by the expected financial return compared with the cost of training. Assuming that people make rational, utility-maximizing decisions, the decision to invest in training will involve estimating the returns over a lifetime and comparing these returns with the costs of the training This model of human capital investment helps explain such questions as why the United States has so many medical specialists compared to general practitioners. The simple answer is that the net financial returns to specialties such as orthopedic surgery are higher than the returns to primary care physician practices. Public policy has attempted to change this somewhat. For example, Medicare's RBRVS method of determining reimbursements to physicians now provides relatively higher payments to primary care physicians than formerly. Nonetheless, a wide difference in incomes of physicians in different specialties, even when adjusted for years of training, still remains. However, over time, the net return to medical training as a whole, compared with many other professions, has declined. This is due in no small part to the growth of managed care in both private and public insurance.

In understanding the market for physician services, it is also necessary to look at the demand side. Demand for medical services has increased with improvements in medical technology, which make it possible to accomplish more improvements in health. This will be discussed further later in this chapter. Medicare (and Medicaid) also brought about a huge increase in the demand for physicians as elderly and low-income Americans could afford more medical care.

This led to a shortage of physicians. Medical schools received more government subsidies. The student loan program for medical training was expanded. Medicare began to

heavily subsidize resident training in hospitals. Immigration and licensing laws were changed to make it easier for internationally trained physicians to immigrate to the United States and to become licensed practitioners once here.

*Registered Nurses*

Training to become a registered nurse (RN) also involves human capital investment, although one that requires fewer years of training. Because of chronic shortages of registered nurses, measured by persistent chronic vacancy rates in hospital nursing positions, there has been a great deal of subsidization of nurse training programs over the years. Although health policy planners still find a shortage of nurses, the nurse/physician ratio as well as the nurse/population ratio in the United States increased until the mid-1990s. Since then, the enrollment in nurse training programs has been declining. One reason for this is that there are now more professions open to women, including becoming an MD.

The market for nurses has been analyzed as a classic case of monopsony. Monopsony exists when there is monopoly power on the part of employers (hospitals) and an upward-sloping supply curve of nurses. It is manifested by a disequilibrium in the form of a gap between supply and demand at a given wage. There is considerable evidence of monopsony in the nurse market between 1940 and 1960. Whether it still exists today is a bit more problematic.

In the absence of monopsony, one expects a shortage to be resolved by a rise in wages until an equilibrium is reached. If shortages persist and wages do not rise, the logical explanation, in the absence of governmental intervention to control wages, is that there is monopsony. In that case, employers find that the marginal return to raising wages is not equal to the marginal cost of hiring more workers, even when there is a gap between demand and supply, such as persistent vacancy rates in positions for hospital nurses. Note that the supply curve is, to the employer, the *average factor cost curve,* whereas the real cost of hiring more workers is the *marginal factor cost.* The marginal factor cost rises faster than the average factor cost because when additional nurses are hired, wages of those already employed have to be raised to create parity. This is the reality of the workplace where productivity will decline if new hires are paid more than other employees doing comparable work.

## The Physician–Patient Relationship

Although the proportion of total expenditure on medical care devoted to physician payments is less than 25%, physicians are of central importance in the provision of medical care in that they organize and direct the path of treatment. Because of asymmetry of knowledge between patients and physicians, patients delegate authority to physicians. This is a good example of a principal-agent relationship in which there is always the possibility of imperfect agency since physicians can substitute their own welfare for that of the patient. Since professional standards forbid this, we can assume that physicians experience some disutility in behaving as imperfect agents and will therefore only do so if there are offsetting benefits from this behavior, such as enhancement of income.

The way in which physicians are paid for their services will have no effect on their treatment of patients if they are perfect agents. But in the case of imperfect agency, physicians paid on a fee-for-service basis may be tempted to recommend greater treatment intensity than is in the patient's best interest. Thus, there may be "physician-induced demand." On the other hand, if payment is on a capitation basis, where physicians agree to treat patients in return for a fixed fee per year, they may be tempted to skimp on the amount of treatment offered in order to handle a larger patient load. A risk-averse physician, worried about the possibility of accusations of malpractice, might also order unnecessary tests. Note that all of these examples apply only to an established patient-physician relationship. If a physician locates an office practice in a wealthy neighborhood in order to attract patients who will pay higher fees, this is not considered imperfect agency, although it may be done to enhance income.

Conventional models treat independent physician practices as monopolistically competitive firms with downward-sloping demand curves since physicians, even those practicing in the same subspecialty, are not perfect substitutes for each other and may have considerable market power based on reputation. A newer model, developed by Thomas McGuire (2000), applies particularly well to a post-managed-care world. In this model, physicians may not be able to set the price, but they can vary the quantity of service provided. McGuire substitutes the notion of a net benefit function for a demand curve. In this framework of analysis, there are substitutes (though imperfect ones) for physicians, and patients will only remain under the care of a given physician if he or she provides a service (net benefit) equal to or greater than some minimum level. Since patients have imperfect knowledge, they may want less treatment than a caring and conscientious physician thinks optimal. So, a patient might leave a physician, even if he or she were behaving as a perfect agent. However, this model can also explain the limits of either physician-induced demand or skimping on service that a patient will permit, in the case of a physician who is an imperfect agent.

## Hospitals

Hospitals are complex organizations. The modern acute-care hospital is a multiproduct firm providing a variety of different in- and outpatient services. The most common form of hospital in the United States is the private nonprofit community hospital, although there are also public (government) hospitals and for-profit private hospitals. Public hospitals tend to have a higher proportion of patients of lower socioeconomic status, and most elite teaching hospitals, associated with medical schools, are private, not-for-profit institutions.

Theories of hospital management are derived from theories of the modern corporation, in which the function of manager and owners is usually separated. They focus on

the behavior of the decision makers and differentiate between hospitals in which the CEOs are nonmedical professional managers (Newhouse, 1970), hospitals that are run by physicians (Pauly & Redisch, 1973), and those in which there is shared management by business managers and physicians who may have differing objectives (Harris, 1977). In all cases, the models assume that managers behave so as to maximize their utility.

In theory, managers of nonprofit organizations would be expected to emphasize quality over quantity or cost minimization compared with managers of for-profit hospitals since they do not distribute profits to owners. However, empirical research finds little difference between for- and not-for-profit hospitals. Studies that compare different ownership types of acute-care hospitals find that there is not much difference, on average, in prices charged, intensity of care, or patient outcomes. With respect to nursing homes and psychiatric hospitals, for-profit institutions are more prevalent but also may provide lower quality care.

Many communities have only one or two hospitals serving the area. Hospitals may be viewed as monopoly firms, particularly when we consider their relationships with employees. Hospitals may in some situations be natural monopolies. A natural monopoly is a firm characterized by long-run economies of scale (a downward-sloping cost curve) over the whole range of its demand. In this case, adding additional firms within the same market will result in higher costs (and prices). However, more commonly we use oligopoly models to analyze the behavior of hospitals as sellers of services since there is usually some degree of competition within a region, and natural monopoly characteristics do not seem to pertain beyond certain capacity levels.

Oligopolies often compete on some basis other than price. This was certainly true of hospitals, at least until the 1990s. This led to the notion of the "medical arms race," where hospital managers compete to have the best facilities and equipment. When one hospital acquires some new technology or opens a new department, other hospitals in the region are forced to follow suit. In this case, competition in a region leads to higher prices and duplication of facilities. The belief that hospital competition is not in the public interest led to the passage of certificate of need (CON) laws. CON laws require government approval to add facilities. The study of effects of CON laws has found some evidence that CON laws provide barriers to entry and lead to higher prices (Salkever, 2000). Moreover, in the post-managed-care era, there is evidence that hospitals do engage in price competition, faced with cost-conscious insurers who themselves have a good deal of market power.

Hospitals are well known to engage in price discrimination. They charge different prices to different third-party payers, public and private, and they may charge lower prices or provide free charity care to the uninsured. The latter is probably based on altruism, although it may be good public relations as well, and nonprofit hospitals are usually required by law to provide a certain amount of charity care. But price discrimination is consistent with a profit-maximizing model of the firm in which higher prices charged to customers whose price elasticity of demand is lower leads to greater profits.

Price discrimination does not necessarily involve cost shifting. The issue of the degree of cost shifting (e.g., charging more to certain consumers *in response to* lowered prices to others) is still not resolved, but there is less evidence of cost shifting than the general public assumes. There is, however, a good deal of cost shifting from individual patients to taxpayers who ultimately pay for much of the subsidized care that hospitals provide.

### Pharmaceuticals

The market for pharmaceuticals is extremely complex. The pharmaceutical industry is heavily regulated, with many countries having some body similar to the U.S. Food and Drug Administration that rules on safety and efficacy of new products. Large drug companies are often thought to be oligopolies, but as the pharmaceutical industry has become global and smaller biotech companies have entered the market, it may be more accurate to think of the pharmaceutical industry as monopolistically competitive. However, drug companies do have some degree of temporary monopoly power in individual product markets when they obtain patents on new drugs. Patent protection is important in providing incentives to innovate. In the drug industry, patents are limited to 20 years, including time when the drug is being developed but is not yet on the market. The pharmaceutical industry, unlike many other industries characterized by rapid technological change (such as computers), is characterized by very high development costs compared with production costs (i.e., the marginal cost of an additional bottle of pills). An implication of this is that consumers in countries that have pharmaceutical industries that innovate are likely to experience higher prices for new drugs since the high cost of discovering the new drug, clinically testing it, and bringing it to market must be recouped, and only a small fraction of new innovations turn out to be marketable at a profit.

There is also a great deal of price discrimination, with cross-country differences in price, within-country price differences between brand-name and generic versions of drugs, and different prices charged depending on the negotiating power of third-party payers. Although on-patent brand-name drug prices tend to be higher in the United States than in many other countries, generic versions of drugs are often cheaper.

## Comparative Health Care Systems: Brief Overviews

### Canada

At approximately the same time that Medicare and Medicaid were enacted in the United States, Canada adopted a universal social health insurance system that covers most medical care for all citizens and permanent residents. Also called Medicare, it is a "single-payer" system in

which the government acts as insurer. It is financed out of tax revenues. There is a separate Medicare budget for each province, jointly funded by the federal and provincial governments, but health insurance is portable throughout Canada. The medical care system is similar to that in the United States in that physicians are reimbursed on a fee-for-service basis and patients are free to choose their own doctors and are not subject to a gatekeeper system.

However, in Canada, global budgets determine how much health care is available, and physicians and hospital associations have to negotiate with the government, which sets rates of remuneration for providers. Technology is less widely diffused. For instance, there are many fewer magnetic resonance imaging (MRI) machines per 10,000 population.

In Canada, there is no option to the public system. Services that are covered by Medicare cannot be purchased privately. Supplementary private insurance can be used only to pay for services not covered by Medicare or to pay co-payments charged by Medicare.

### United Kingdom

After World War II, the National Health Service (NHS) was enacted in the United Kingdom. It differs from the Canadian system in that it is a national health system, not a nationwide universal insurance system. Doctors who participate in the NHS are employees of the government and are paid by a mix of salary and capitation. Originally, hospitals contracted directly with the district health authorities who paid them for their services. In the early 1990s, the system was altered to give hospitals and physician practices some degree of autonomy. Hospitals are now often organized as trusts. Large regional groups of physicians are given their own budgets to manage. However, the source of funding is still the government, and therefore global budget caps apply. The United Kingdom is well known for high-quality care but long waits for all but emergency services, known as "rationing by queuing." In the United Kingdom, unlike Canada, there is the option to "go private" and purchase services outside the NHS system. These services are paid for out of pocket or by private insurance. Physicians in private practice are paid on a fee-for-service basis.

### Germany

Germany has a system of many competing "sickness funds" that are nonprofit insurers. Most workers join sickness funds through their place of employment, though unions and other community organizations also provide access to the funds. The sickness fund system is social insurance in that premiums are financed through a payroll tax, which, like the Medicare Social Security tax in the United States, is jointly paid by workers and employers. It is, however, much higher than in the United States. Retired persons and the self-employed are also enrolled in sickness funds and contribute in proportion to their pensions or self-employment earnings. Only the very wealthy may opt out of this social insurance system. About 95% of all German citizens are members of the sickness funds.

Since the sickness fund system is a not a single-payer system, there is the problem of adverse selection. To offset this, government regulation requires sickness funds to cross-subsidize each other so that those that have a more expensive pool of members receive payments from other funds whose expenses are lower.

Physicians associations negotiate fees with the sickness funds served by their members. The German government has imposed global caps on different components of the medical budget. If a group of physicians serving a particular sickness fund exceeds their budget cap, every member of the physician group is subject to a reduction in the rate of remuneration (fee schedule).

Reforms in the German system have increased co-payments and have introduced a degree of competition in that individuals or worker groups may choose which sickness fund to join and may shift their membership if they are dissatisfied. The German system has, historically, been very generous in its coverage of services. Physician visits are still free, although some fees are now charged for prescription drugs, eyeglasses, and so on.

In each of these countries, demand for medical services is rising, and the public systems, all of which have global budgets, are finding their finances strained. Services have to be rationed, by queuing up for nonemergency care, by charging fees for formerly free services, or by denying certain kinds of treatment. In Germany and the United Kingdom, it is possible to purchase services through the private market. In Canada, this can be done only by crossing the border and purchasing medical care in the United States.

### Public Health Care in the Developing World

Although it is difficult to generalize about low-income nations in different parts of the world, there are certain common characteristics. Communicable diseases represent a higher proportion of the populations' sickness. A smaller proportion of the GDP is generally devoted to health care. Rural areas are often disproportionately lacking in health care facilities. And even countries that have recently experienced dramatic rates of economic growth, such as India and China, do not provide adequate free public health care, even to the poor. International agencies such as UNICEF and nongovernmental organizations (NGOs) play an important role in providing health care and medicine to the developing nations. This is particularly important in Africa, given the high incidence of HIV/AIDS.

## Leading Proposals for Reform in the United States

The proportion of national income spent on health care in the United States is greater than in any other industrialized country. It is now in excess of 15%. Our medical technology, measured in terms of such indices as number of MRI units

per 10,000 population, also exceeds that of other comparable countries. Yet, aggregate statistics (averages) of health outcomes place the United States far from the top of a list of comparable (Organisation for Economic Co-operation and Development) nations in life expectancy, both at birth and at 60 years of age, and U.S. infant mortality rates are discouragingly high. The United States is also the only high-income industrialized nation that does not have universal health insurance coverage for its citizens and permanent residents. More than 15% of families were without health insurance coverage in 2009. The fact that health insurance and medical care costs are increasing much more rapidly than the consumer price index, although not unique to the United States, also leads to the conclusion that the system needs reforms both with respect to its efficiency and its equity.

## Incremental Reforms

Improvements in efficiency could be defined as those that result in the same quality of care provided at lower cost and/or better patient outcomes provided at no higher cost.

### *Use of Information Technology to Provide Better Records*

Using information technology (IT) to provide comprehensive patient records would both reduce medical errors, such as prescribing drugs that are counterproductive, given patient allergies or when combined with other medications, and provide cost-savings by eliminating unnecessary duplication of diagnostic tests.

Another use of IT is the expansion of health care provider report cards. Although critics fear that providers will attempt to avoid treating the most difficult cases that might spoil their records, risk/adjustment applied to reporting can largely overcome this difficulty, and public support for making such information available is now widespread.

The use of IT to simplify and standardize insurance claim forms is also broadly advocated. A related proposal requiring that all health insurance contracts cover certain basic services is more controversial. Critics argue that this would curtail freedom of choice if "consumer-driven" insurance options were ruled out. However, proponents of regulation believe that most consumers would benefit from requirements that all plans cover a broad range of medically necessary treatments and not exclude any on the basis of preexisting health conditions.

### *Better Management of Care for Chronic Diseases*

Another widely accepted reform is better ongoing care for patients with chronic illnesses. This would be facilitated by more continuity in health care provision as well as better recordkeeping over the lifetime of patients. Critics of employment-based health insurance see it as a stumbling block to long-term management of chronic conditions in a world in which there is so much mobility between jobs and in which employees are subject to employers' decisions to change insurance carriers based on cost considerations. Proposals for reforms that involve greater portability of health insurance address this problem. Advocates of a "single-payer" universal insurance system believe that this and other inefficiencies associated with the fragmented insurance system in the United States would be best overcome by having the government act as third-party payer for all citizens and residents.

### *Promoting Wellness*

A third widely accepted reform is a health care system that provides incentives for a healthier lifestyle. Advocates favor promoting, through subsidies, types of preventive care that have been shown to be effective. This includes both screening for disease and programs that promote a healthier lifestyle.

### *Reforming the Tax Treatment of Employment-Based Health Insurance*

This has been advocated for a number of decades by many economists who regard treating all employment-based health insurance premiums as tax-free income to be both inefficient and inequitable in that it encourages workers to demand excessive insurance coverage, which in turn promotes cost insensitivity and the use of health care services with low marginal value. It also benefits high-income workers disproportionately since they benefit more from the tax subsidy. Proposed reforms include capping the level of health insurance premiums that will receive favorable tax treatment and completely removing the tax-free status of insurance premiums.

### *Broadening Insurance Coverage*

Although most people who advocate universal health insurance coverage in the United States are primarily concerned with equity or fairness, there are also inefficiencies associated with having approximately 47 million people currently uninsured. These include the inappropriate use of hospital emergency rooms by the uninsured who have no access to physician office visits and the postponing of treatment until advanced stages of disease. However, covering the uninsured would not be, at least in the short run, cost free (Institute of Medicine, 2003).

## The Access Problem

There is a widespread belief in the United States that all residents should have access to affordable health insurance, but there is no general agreement on the best way to achieve this goal. Several main proposals are espoused by health economists, although the details differ. Some involve incremental change, building on our combination of existing employment-based private insurance and social insurance programs. Other plans would provide more

dramatic changes by replacing employment-based health insurance altogether. Victor Fuchs and Ezekiel Emanuel (2005) have grouped reforms of the health care system into three main categories: (1) incremental reforms, such as expanding SCHIP or expanding Medicare to cover 55- to 65-year-olds, or individual or employer mandates, which create new insurance exchanges and provide subsidies to low-income families but do not radically alter the structure of the current health care system; (2) single-payer plans that would eliminate the private insurance market and have the government act as third-party payer; and (3) voucher-based reforms, which they advocate (Furman, 2008, chap. 4). Although they do not explicitly consider health savings accounts (HSAs), this is a fourth option that we need to include in a complete menu of proposed reforms.

1a. *Expand Medicare.* This can be accomplished by allowing people to buy into the Medicare program. One way to do this would be to allow anyone to pay a standard premium and buy Medicare as an alternative to private insurance. More ambitious proposals would gradually phase out Medicaid and other forms of social insurance, subsidizing low-income families' Medicare premiums and reducing the co-payments for them. There might, however, have to be some kind of subsidy from the federal government if Medicare suffered from adverse selection, compared with employment-based private insurance.

1b. *Mandate Individual Health Insurance Coverage: The Massachusetts Plan as Blueprint.* Massachusetts has instituted a plan for statewide universal health insurance coverage, achieved through a mandate that individuals must have health insurance or pay a fine. This plan involves a combination of private and public health insurance coverage, with Medicare, Medicaid, and SCHIP remaining in place. Low-income families that are not currently covered by social programs receive subsidies to cover all or part of their health insurance premiums. Most workers who currently have employment-based insurance are expected in the short run to remain with these plans. However, they have the option to acquire health insurance through the Commonwealth Connector, a market clearinghouse through which insurance providers can offer portable health plans that function like employment-based plans of large employers. To make the system work, the portable plans must be subject to the same tax treatment as employer-based plans. Other states are considering similar plans. Jonathan Gruber (2000) is the architect of a national health insurance system based on Massachusetts's plan.

2. *Single-Payer System.* Universal health insurance in the form of a single-payer system, modeled on Canadian Medicare, has been proposed (e.g., Rice, 1998) but has generally not been considered politically feasible, although it has had the support of some presidential candidates and members of Congress. Even though the U.S. Medicare program is more cost-efficient than most private health insurance, and the simplicity of a Canadian-type single-payer system is appealing and removes the problem of adverse selection, there is widespread belief in the United States that government- run programs have no built-in mechanisms that ensure efficiency and are likely to be subject to corruption. Moreover, Medicare itself in its present form is no longer a single-payer system since Parts B and D allow beneficiaries to assign their benefits to private insurers.

3. *A Voucher System* (Fuchs & Emanuel, 2005). All U.S. residents would receive a health care voucher that would cover the cost of an insurance plan with standard benefits. In the short run, those who are currently insured through Medicare, Medicaid, SCHIP, and so on would have the option of remaining in those plans or switching to the voucher plan. However, the existing forms of social insurance would gradually be phased out, and employment-based insurance would be discontinued. The latter would lose its attractiveness since part of the reform plan is to discontinue the favorable tax treatment of employment-based group insurance.

The voucher has no cash value but gives the recipient the right to enroll in a health plan with standard benefits. National and regional health boards, modeled on the Federal Reserve System, would regulate insurance plans with respect to both their finances and their provision of adequate networks of providers. Although the system would be universal, third-party payers would be private. People could choose insurance programs or, if they failed to do so, be assigned to one. The system assumes that private insurance companies would have an incentive to remain in the market and compete for subscribers. In that sense, it is not unlike the managed competition model of health insurance markets (Enthoven, 1993). The Fuchs and Emanuel (2005) plan would be financed by a value-added tax on consumption, but a voucher plan could also be financed out of an income tax.

4. *Health Savings Accounts (HSAs).* HSAs, whose advocates often structure the plans as forms of tax-free income, are not insurance per se. They are personal savings accounts, similar to 401K or 403B retirement accounts, and employers could contribute to them instead of supporting group health insurance. One of the earliest advocates of HSAs was Martin Feldstein (1971). An advantage of HSAs is that they encourage judicious use of health care since individuals are paying the bills themselves out of their own savings. A disadvantage is the loss of pooling of risk across individuals. What is retained is the individual's or family's ability to smooth expenditure on medical care over periods of health and illness. For such plans to constitute anything close to a universal system, low-income families would need to have their HSAs heavily subsidized. HSAs could be limited to medical care after retirement, in which case they would be an alternative to Medicare, or they could replace employment-based

health insurance as they have in Singapore. Singapore's HSAs are supplemented by social health insurance for catastrophic health expenses.

## Conclusion

The United States is widely acknowledged to need health care reform. More than 47 million people without insurance is considered unacceptable by most people. Changes in the nature of the labor market have made employment-based health insurance less appropriate than it was when long-term employment with the same firm was usual. It provides job lock and reduces the competitiveness of U.S. firms. The proportion of the GDP devoted to health care is much higher than in other countries without better patient outcomes. A well-conceived reform plan could significantly lower health care costs and provide much greater equity (access). However, technological advances in medical care and demographic trends will almost inevitably lead to a continued upward trend in the proportion of the budget devoted to medical care (Newhouse, 1992). This is, however, not unique to the United States but is a problem facing all industrialized nations.

## References and Further Readings


Arrow, K. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review, 53,* 941–973.

Berndt, E. R. (2002). Pharmaceuticals in U.S. health care: Determinants of quantity and price. *Journal of Economic Perspectives, 26,* 45–66.

Currie, J., & Gruber, J. (1996). Health insurance eligibility, utilization of medical care, and child health. *Quarterly Journal of Economics, 111,* 431–466.

Cutler, D., & McClellan, M. (2001). Is technical change in medical care worth it? *Health Affairs, 20,* 11–29.

Cutler, D., & Zeckhauser, R. K. (2000). The anatomy of health insurance. In A. J. Culyer & J. P. Newhouse (Eds.), *Handbook of health economics* (Vol. 1A, pp. 563–643). Amsterdam: Elsevier.

Enthoven, A. C. (1993). The history and principles of managed competition. *Health Affairs, 10,* 24–48.

Feldstein, M. S. (1971). A new approach to national health insurance. *Public Interest, 23,* 93–105.

Finkelstein, A. (2007). The aggregate effects of health insurance: Evidence from the introduction of Medicare. *Quarterly Journal of Economics, 122,* 1–37.

Fuchs, V. R., & Emanuel, E. J. (2005). Health care vouchers: A proposal for universal coverage. *New England Journal of Medicine, 352,* 1255–1260.

Furman, J. (Ed.). (2008). *Who has the cure? Hamilton Project ideas on health care.* Washington, DC: Brookings Institution Press.

Glied, S. A. (2000). Managed care. In A. J. Culyer & J. P. Newhouse (Eds.), *Handbook of health economics* (Vol. 1A, pp. 707–753). Amsterdam: Elsevier.

Glied, S. A. (2001). Health insurance and market failure since Arrow. *Journal of Health Economics, 26,* 957–965.

Grossman, M. (1972). On the concept of health capital and the demand for health. *Journal of Political Economy, 80,* 223–255.

Gruber, J. (2000). Health insurance and the labor market. In A. J. Culyer & J. P. Newhouse (Eds.), *Handbook of health economics* (Vol. 1A, pp. 645–706). Amsterdam: Elsevier.

Gruber, J. (2008). Taking Massachusetts national: Incremental universalism for the United States. In J. Furman (Ed.), *Who has the cure? Hamilton Project ideas on health care* (pp. 121–141). Washington, DC: Brookings Institution Press.

Harris, J. E. (1977). The internal organization of hospitals: Some economic implications. *Bell Journal of Economics, 8,* 467–482.

Institute of Medicine. (2003). *Hidden costs, value lost: Uninsurance in America.* Washington, DC: National Academy of Sciences.

Johnson-Lans, S. (2006). *A health economics primer.* Boston: Pearson/Addison-Wesley.

Kremer, M. (2002). Pharmaceuticals and the developing world. *Journal of Economic Perspectives, 16,* 67–90.

McGuire, T. G. (2000). Physician agency. In A. J. Culyer & J. P. Newhouse (Eds.), *Handbook of health economics* (Vol. 1A, pp. 461–536). Amsterdam: Elsevier.

Newhouse, J. P. (1970). Toward a theory of nonprofit institutions: An economic model of a hospital. *American Economic Review, 60,* 64–74.

Newhouse, J. P. (1992). Medical care costs: How much welfare loss*? Journal of Economic Perspectives, 6,* 3–22.

Newhouse, J. P., & the Insurance Experiment Group. (1993). *Free for All? Lessons from the RAND Health Insurance Experiment.* Cambridge, MA: Harvard University Press.

Pauly, M. V. (2001). Making sense of a complex system: Empirical studies of employment-based health insurance. *International Journal of Health Care Finance and Economics, 1,* 333–339.

Pauly, M. V., & Redisch, M. (1973). The not-for-profit hospital as a physicians' cooperative. *American Economic Review, 63,* 87–100.

Reinhardt, U. E. (1972). A production function for physician's services. *Review of Economics and Statistics, 54,* 55–65.

Rice, T. (1998). Can markets give us the health system we want? In M. A. Peterson (Ed.), *Healthy markets: The new competition in health care* (pp. 61–103). Durham, NC: Duke University Press.

Salkever, D. S. (2000). Regulation of prices and investment in hospitals in the United States. In A. J. Culyer & J. P. Newhouse (Eds.), *Handbook of health economics* (Vol. 1B, pp. 1489–1535). Amsterdam: Elsevier.

# 70

# Economics of Health Insurance

Amy M. Wolaver
*Bucknell University*

Despite the fact that the United States spends more money per capita than any other nation, not everyone is insured. Perhaps even more surprising, when one adds the various government indirect subsidizations of health care through the tax code to the direct expenditures on Medicare, Medicaid, and other programs, the total dollars spent per capita by the *government* on health care is also higher than any other industrialized nation, in which all achieve universal coverage (Woolhandler & Himmelstein, 2002).

Because financing of health care in the United States is achieved through a mix of public and private insurance, the system is complex. Health insurance markets are riddled with market failures, which have important implications for who is covered and how coverage is achieved. Other industrialized countries are able to achieve universal coverage, even in systems that also mix public and private coverage. What are the market problems in health insurance that lead to gaps in coverage we experience in the United States?

This chapter will focus on the economics underpinning the market for health insurance in the United States. The primary goal of the chapter is to understand the impact of the market failures in the financing of care. First we will model demand for insurance and outline the market failures in insurance. We will examine some of the market and policy solutions to these market failures. We will briefly examine the history of health insurance coverage in the United States and how tax policy affects where Americans obtain their coverage. We will outline some of the major features of the forms of publicly provided health insurance, Medicare and Medicaid/State Child Health Insurance Program (SCHIP). We will examine the characteristics of the uninsured population in the United States in relation to the institutions and market failures outlined in the chapter. We conclude with a case study on HIV/AIDS to illustrate some of the complexities in achieving health insurance coverage.

## The Demand for Health Insurance

The perfectly competitive model in economics assumes perfect information, but in reality, there are many gaps in our information set. Kenneth Arrow's (1963) article, the cornerstone of health economics, beautifully describes the problem of uncertainty. We cannot know the future, and we face the possibility of loss of income and assets from health problems. The loss of income and assets comes through lost work time and from medical bills. At the same time, most of us would prefer to not face this risk of loss. This feature of our preferences, risk aversion, arises because most people have decreasing marginal utility in wealth or income. That is, the increase in utility from the first dollar of income you receive is higher than the increase in utility from your hundredth dollar and so on. Because of this feature, we dislike losing a dollar more than we like gaining a dollar. In economic terms, the decrease in utility from losing a dollar is bigger than the increase in utility from a dollar.

Economists model planning in the face of uncertainty by using probabilities and expected values. Suppose you know you have a 10% chance of having a health problem that will cost you $10,000, and normally your income is $50,000. Therefore, 90% of the time, your income is $50,000, and 10% of the time, it is $40,000. Your expected income ($E$[income]) for the year is therefore

$$E[\text{income}] = 0.90 * \$50{,}000 + 0.10 * \$40{,}000 = \$49{,}000.$$

The expected loss in this case is the probability of loss times the amount of the loss. In this case, $E$[loss] = 0.10 * \$10,000 = \$1,000.

However, economists believe that it is expected utility that matters when you are planning for the future, not expected income. Because of risk aversion, expected utility with that uncertain 10% chance of a \$10,000 loss is different from the utility one would have with a 100% chance of a loss of \$1,000, even though the losses in expected income are equivalent. Let us assume that we can characterize the utility function as $U = \sqrt{I}$, where $U$ is utility and $I$ is income. This is a simple function that is consistent with risk aversion; in other words, there is a decreasing marginal utility of income. Table 70.1 shows the level of utility for each of the possible levels of income with this function.

Expected utility with the 10% chance of loss is

$$E[\text{utility}] = 0.10 * 200 + 0.90 * 223.6 = 221.24.$$

Compare the expected utility with an uncertain loss to the utility with certain income of \$49,000. Utility in the latter case is 221.4, higher than the expectation with loss. We will not place any direct interpretation on what one unit of utility means; the important feature is that you would have higher utility losing \$1,000 every year with 100% certainty than you have with a 10% chance of losing \$10,000 each year. This decrease in utility is risk aversion. Because of risk aversion, consumers are willing to pay someone to take away some of the financial uncertainty; this product is insurance. The discomfort from risk aversion increases with the level of uncertainty (risks that are closer to 50/50 rather than 0% or 100% risks). The discomfort also increases with the possible loss. Insurance has therefore traditionally been more likely to cover high ticket expenses that are uncertain.

### What Is Insurance?

Given that we dislike risk, the following question then arises: Is there a market to pay someone to take the risk for us? The answer is yes; this market is insurance. Essentially, insurance is a product where consumers pay a price, the premium, to some other entity, the insurer, who then assumes the financial risks. After the policy is purchased, if the consumer in our example has a lucky year, with no loss, the income available to spend on other things or to save is \$50,000 minus the premium. In an unlucky year, where the consumer becomes ill and has \$10,000 in medical bills, the insurance company pays the bills for the consumer. Disposable income remains \$50,000 minus the premium for the insurance.

Assume for simplicity's sake that there is an entire population of people just like our hypothetical consumer, with a 10% chance of a loss of \$10,000. From the insurer's perspective, if the premium they charge is \$1,000, and they sell 10 policies, on average, they will collect \$10,000 in premiums and pay out one claim of \$10,000, and they break even. However, this scenario ignores the administrative costs of running an insurance company. The company incurs administrative costs in selling the policies, in paying claims, and in designing their policies. Given these costs, to make normal (zero economic) profits, the insurance company must charge a premium equal to the expected loss plus the administrative costs, also known as a loading factor.

Because of the gap in the expected utility of taking our chances with a loss and the expected utility with a constant loss, consumers are willing to pay a loading factor, up to a certain point. However, the analysis becomes more complicated when we realize that different people face different chances of illness and medical-related losses.

**Table 70.1** Utility at Different Incomes

| *Income* | *Utility* |
|---|---|
| \$40,000 | 200 |
| \$49,000 | 221.4 |
| \$50,000 | 223.6 |

### Types of Insurance

The market for insurance has evolved into many different forms. The traditional form of insurance is fee-for-service, where health care providers are reimbursed for each service performed. Because of the fast growth in expenses in the health care sector, insurers have experimented with other forms of insurance. Managed care organizations (MCOs) are one market response to this growth in expenses. These organizations may take many forms, including health maintenance organizations (HMOs), preferred provider organizations (PPOs), and point-of-service (POS) plans. Cost savings are achieved through a variety of mechanisms. One mechanism is the use of monopsony (buyer) power to lower fees paid to the providers (hospitals, doctors, pharmacies, etc.). Because the insurer represents multiple potential patients, they may have market power to achieve price discounts. Another is capitated (prepayment) reimbursement mechanisms, which give incentives to providers to reduce the amount of care given. Other mechanisms are to emphasize preventive care and make use of information technologies to reduce the more expensive hospital care.

## Market Failures in Insurance

To introduce the two main market failures in insurance, let us begin with an anecdote. Steve has a Mustang convertible, and the radio has been stolen. While shopping for a

new radio, he finds a store with a 100% theft guarantee; if someone steals the radio, the store will replace it for free. Steve happily purchases the radio and begins to leave the top down on the car when he parks. After the third time the new radio is stolen, the store refunds his money and refuses to replace it again.

Steve's story illustrates the two major problems facing both profit-maximizing insurance companies and the efficient working of the market: adverse selection and moral hazard. The replacement guarantee on the radio is a form of insurance; Steve no longer had to worry about the risk of loss. Since his radio had been stolen once before, Steve knew he had a high risk of theft, and therefore this insurance had more value to him than many other consumers. However, Steve was not one of the customers the electronics store would have preferred; they would prefer customers with a low risk of theft. The first part of the story is an example of adverse selection; the consumers with the highest demand for insurance at any given premium/price are those with the highest risk/expected losses, assuming that income and ability to pay are not an issue.

Once Steve had purchased the radio, his behavior changed. He began to leave the top down on the car because the guaranteed replacement meant he faced lower marginal cost of having his radio stolen. This behavior is indicative of moral hazard; the presence of insurance alters our incentives at the margin.

## Market Failure 1: Adverse Selection

Translating this anecdote to the general issue of health insurance markets, we must first ask, how can adverse selection exist? The underlying problem is one of asymmetric information; one party in the transaction has more information than the other party. In this case, the consumer has more information about his or her risk than the seller (the insurer) does. The insurer cannot know with certainty the expected losses facing any individual, and the individual has an incentive to hide this information to avoid paying a higher premium for coverage. This type of market also has been termed a "lemons" market, by George Akerlof (1970), who coined it from the used car market.

When hidden differences in health risks exist between individuals, insurers have to figure out the profit-maximizing response. To examine this problem further, ignore the administrative costs and risk aversion for a moment and assume that one third of the population has a 15% chance of a $10,000 loss, another third has a 10% chance of the same loss, and the final third of the population has a 5% chance of loss. If an insurer offers a policy to cover the $10,000 loss based on the average population risk, the premium will be

$$\text{Average population loss} = (0.05 * \$10{,}000 + 0.10 * \$10{,}000 + 0.15 * \$10{,}000)/3 = \$1{,}000.$$

A consumer compares his or her expected losses without insurance to the premium he or she must pay to avoid the losses. If the premium is less than or equal to the consumer's expected losses (remember, we are ignoring risk aversion at the moment), he or she will purchase the insurance; otherwise, he or she will not. Consumers with the 5% loss have an expected loss of only $500, so they will not purchase insurance, but the consumers with the two higher levels of risk will.

The insurance company will have negative profits. Half of their policies were sold to consumers with a 15% risk and half to consumers with a 10% risk. They collect $1,000 in premiums per person but make average expected payouts of $(0.15 * \$10{,}000 + 0.10 * \$10{,}000) \div 2 = \$1{,}250$, for a per person loss of $250. If we imagine the company executives do not realize the problem of adverse selection and analyze their first profits, they might assume that the problem was that they miscalculated the population risk initially and feel they have better data now. In the following period, they might charge $1,250 in premiums. However, since consumers behave rationally, only those with a 15% chance of the loss will purchase the policies at the new price, resulting in still more losses for the insurer. As the cycle continues, the insurer will continue to raise premiums and insure fewer people, until the market itself might disappear. This process of increasing premiums and decreasing coverage has been called a death spiral and in extreme circumstances could lead to the disappearance of the entire market. Because of risk aversion, the lack of a market for insurance will decrease welfare (Arrow, 1963).

### *Solutions to Adverse Selection*

Of course, insurers are not as naive as our above scenario would suggest. They are well aware of the problem of asymmetric information and the resulting adverse selection. The source of the death spiral is the free entry and exit of consumers in this market, combined with their hidden information. While you as a consumer do not have perfect information about your health risks, you do have more information about your own behaviors, family history, and so on than an insurer does. Because consumers are utility maximizing, the consumers who find it worthwhile to purchase insurance are the relatively high-risk consumers in this free entry and exit model. Insurers, consumers, and the public have developed mechanisms to at least partially solve the problems caused by adverse selection. We explore several of the most important solutions below, although this list is not exhaustive.

### *Market Solution 1: Group Insurance and Risk Pooling*

If insurers could find a way to have a pool of consumers with both high and low risks, they could continue to make normal profits, and everyone would have access to insurance. Low-risk individuals would pay a higher premium than their

average expected losses, and high-risk individuals would pay a lower premium than their average expected losses.

This risk pooling mechanism is essentially group insurance. In the United States, the grouping mechanism traditionally has been employers, for reasons we will explore more fully below. If we now remember that there are administrative costs to add to the premium, we can find one additional advantage above that of risk pooling to group insurance. Administrative costs decrease rapidly with the size of the group. If a salesperson contracts to insure a group of 250 employees, he or she has, in essence, sold 250 policies in one fell swoop, a less expensive alternative to selling 250 separate policies to 250 consumers. The employer offers the insurance to all employees as a work benefit.

Why would a low-risk individual be willing to, in effect, subsidize a high-risk individual through coverage at work? Several reasons exist; first, because of risk aversion, everyone is willing to pay something more than the value of their expected losses. Second, the cost of purchasing a dollar's worth of health insurance at work is less than a dollar because of the tax treatment of these benefits and because of the lower administrative costs. Third, a low-risk individual today might be a high-risk individual tomorrow, and there might be barriers to purchasing insurance later. In this situation, workers will find it to their advantage to obtain coverage early.

### *Market Solution 2: Benefit Design and Risk Segmentation*

Insurers can also offer multiple types of policies designed to make individuals "signal" their risk level by which policy they choose to purchase. In essence, insurers separate risks and price the policies appropriate to the risk level. They also make use of observable characteristics correlated with health care expenses in setting the availability of the product. To discuss these options, let us first define several characteristics or terms related to health insurance coverage.

The premium is the price the consumer pays in exchange for the insurance coverage. If the consumer has a loss, he or she makes a claim and is reimbursed, or providers are paid directly. Deductibles are exemptions from reimbursement for the first dollars of losses. As an example, someone with a $500 deductible and a $750 medical bill will pay the first $500 out-of-pocket (himself or herself), and the insurance company will pay the next $250. Coinsurance is an arrangement where the insurance company pays a certain percentage of the claims and the consumer pays the remaining percentage out-of-pocket. A co-payment is either a fixed-dollar payment made by the consumer or the actual out-of-pocket payment by the consumer resulting from a coinsurance arrangement.

As an example of how benefit design can separate risks, insurers may increase the premiums but also increase the level of benefits by decreasing the deductibles, co-payments, and coinsurance rates. Compared to a policy with a low premium but high deductibles and co-payments, the premiums of both policies can be designed jointly with the benefits so that relatively high-risk individuals will find one policy more attractive and relatively low-risk individuals will find the other policy more attractive.

Insurers also spend considerable effort identifying observable characteristics (age, gender, etc.) that are correlated with higher risks and pricing the policies accordingly, as young, unmarried males purchasing automobile insurance can certainly attest. These factors include past medical conditions and other risk behaviors such as smoking. In addition to charging different premiums to different groups, insurers have written in a number of preexisting conditions clauses that exempt coverage from medical bills related to health problems that have already been diagnosed.

All of these efforts at risk segmentation add considerable costs to the loading factor associated with health insurance. These costs are one reason why the administrative costs for private insurance in the United States are considerably higher than the administrative costs in Canada, which has universal coverage (Woolhandler, Campbell, & Himmelstein, 2003). These costs are partly incurred by the insurers themselves in their benefit design and pricing mechanisms, but they also add administrative costs to the providers, who must deal with multiple payers with different reimbursement mechanisms.

### *Policy Solutions to Adverse Selection*

The underlying problem with asymmetric information is that healthy consumers opt out of policies, leaving sicker consumers in the pool. Insurers are wary of the very consumers who are most likely to want to purchase insurance. If everyone in the community is covered, however, and opting out can be prevented, then the problem of adverse selection is alleviated. Several policies can be used to address this problem. We will focus on three: individual mandates, employer mandates, and single-payer plans. The recent extensive health care reforms in Massachusetts rely in part on the use of both types of mandates, although the subsidization of risk pools for individuals and small groups is another feature of this plan.

Mandates essentially require individuals to purchase insurance and/or employers to offer insurance to their workforce. In this case, everyone is now forced to demand insurance, and the willingness to purchase insurance is no longer a signal of hidden high risk to the insurer. Single payer, where the government is the insurer, is an option where the entire population is in one risk pool.

### *Policy Solution 1: Individual Mandates*

Minimum requirements for automobile insurance are one example of individual mandates. Car insurance differs from mandated health insurance, however, in that if one

cannot afford to purchase the policy, one can simply not drive. Health insurance mandates, in contrast, require everyone to purchase a policy. Unlike driving, one cannot simply forego medical treatment in all cases because emergency rooms are required to give treatment in cases of potentially life-threatening health problems. Thus, policy proposals that include mandates typically also include some form of subsidization of coverage through the tax code and/or increased access to public insurance as part of the plan. Creating or fostering risk pools outside of the existing group market is another feature that enhances the success of individual mandates. Penalties for not purchasing insurance under a mandate plan are also typically enacted through the income tax code. But because everyone is required to purchase insurance, low-risk individuals cannot opt out of the pool. The desire to purchase insurance is no longer an indicator of adverse selection.

*Policy Solution 2: Employer Mandates*

Employer mandates differ from individual mandates in several ways. First, employers are typically required to offer insurance, but individuals are not necessarily required to accept, or "take up," that offer. Individuals may have alternative access to insurance through a spouse, for example. Firms that do not offer coverage are expected to pay a financial penalty, which is then used by the government to subsidize public provision of health insurance to individuals without access to employer-provided insurance. By forcing the expansion of the group market, more consumers will get access to the benefits of risk pooling.

Employer mandates will certainly have additional impacts outside of health care on the market for labor. A simple theoretical analysis of mandates is presented in Summers (1989). Figure 70.1 shows one possible outcome in the labor market of an employer mandate. Requiring employers to provide insurance to their workers adds to the cost of hiring labor, which will reduce demand from $D_{\text{no insurance}}$ to $D_{\text{insurance}}$. However, providing access to insurance at work may also increase the supply of labor from $S_{\text{no insurance}}$ to $S_{\text{insurance}}$. The overall impact on the market will be to lower wages to cover some fraction of the added costs of insurance.

As in any market where there is a simultaneous decrease in demand and increase in supply, the impact on the equilibrium number of workers hired is ambiguous. The new equilibrium could fall anywhere in the shaded area. Where exactly the new equilibrium falls depends on whether labor supply increases more, less, or the same amount as the decrease in demand for labor. If supply increases at the same rate as the demand decreases, the number of employed workers would remain unchanged and the wage offset exactly equals the cost of providing insurance. If supply increases *more* than demand decreases, which occurs if workers value the insurance more than it costs the employer to provide the insurance, the equilibrium quantity of workers could even increase. This latter effect could exist because of the benefits in risk pooling through group insurance in combination with the preferential tax treatment of health insurance purchased at work, which is described in more detail below.

*Policy Solution 3: Single-Payer Financing*

Even with mandates, adverse selection problems may remain, however, in the sorting of consumers and groups into different policies. Some insurers could attract relatively high-risk pools while others low-risk pools, reducing some of the subsidization of the high-risk individual. Mandates are not a complete answer to the market failure caused by asymmetric information. A single-payer system is one in which the public sector takes on the role of the

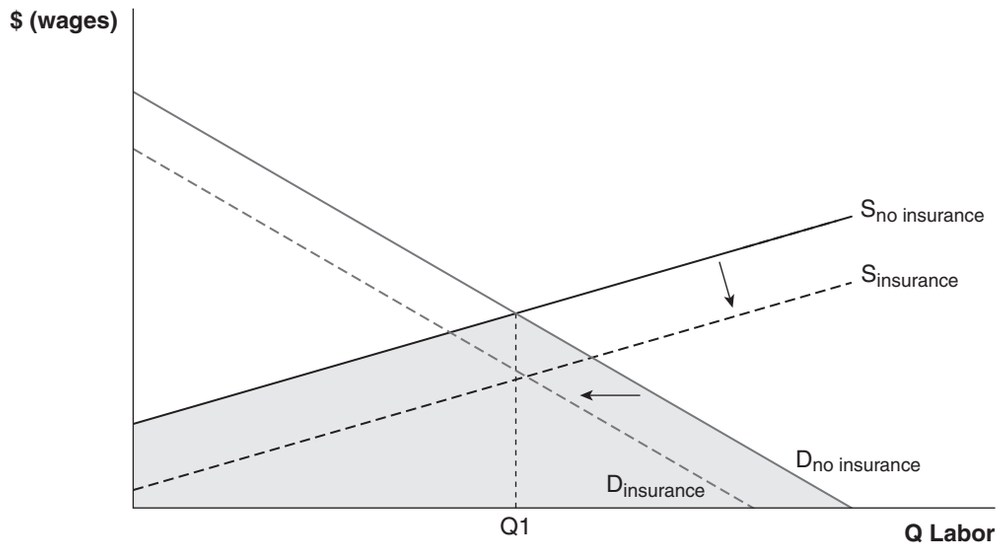


**Figure 70.1** Possible Theoretical Impacts of Mandating Employer-Provided Insurance

only insurer. Financing of the care is typically through income taxes in which the healthy and wealthy subsidize the sick and poor. Universal health coverage is achievable without single-payer financing, but single payer has the advantage of virtually eliminating the problem of adverse selection, depending on how comprehensive the benefits associated with this form of financing are.

The Canadian health care system is one example of single-payer financing that retains private markets for the health care providers. In contrast, in the United Kingdom, the public sector is both the main financer and provider of health care. In the United States, Medicare is a single-payer financing mechanism for the elderly and disabled. In each of these examples, some private financing of health care coverage occurs through out-of-pocket payments by the patients themselves and the existence of supplementary insurance.

## Market Failure 2: Moral Hazard

We now turn to the second major market failure in insurance: moral hazard. Remember Steve's behavior after he purchased the radio? He was less careful about securing the car when he knew the radio would be replaced if stolen. Economists characterize this behavior as moral hazard. Economists do not generally view the problem of moral hazard in health insurance as being primarily related to the idea that people engage in riskier activities because they are insured, although this behavior could be one aspect of moral hazard. Rather, the main problem with moral hazard in health insurance is that once an individual is insured, the marginal private cost of receiving care is reduced from the true marginal resource (social) cost of production of that care. The price facing the insured patient could even be zero; in other cases, the price is reduced from the full cost to a modest co-payment or coinsurance rate.

The efficiency, or welfare, costs of moral hazard are shown in Figure 70.2 for a sample market of doctor visits. To keep things simple, we will assume constant marginal costs, which represent the true, social resource costs of production. In perfectly competitive markets, the equilibrium market price will also be equal to the marginal cost of production. The uninsured, whether their demand curve is represented by $D1$ or $D2$, pay the full marginal costs of care, and the quantity demanded for this group is $Q1$, where the market price ($MC$ curve) intersects the demand curve. This point is the efficient level of production. Again for simplicity's sake, assume for the moment that insurance is full coverage—there are no deductibles or co-payments.

Whether a deadweight loss from moral hazard exists depends on the elasticity of demand for health care. Figure 70.2 shows both scenarios; $D1$ represents the perfectly inelastic demand case, and $D2$ represents an elastic demand curve. Full insurance moves the marginal personal cost (price facing the individual) from the market price to zero. If demand is perfectly inelastic, the quantity demanded of health care remains the same, $Q1$, and the efficient level of consumption is maintained. When demand is elastic, the rational, utility-maximizing individual will optimally consume doctor visits up to $Q0$, where the demand curve intersects the $x$-axis. As is evident, $Q0$ is greater than $Q1$, and the individual is consuming an inefficiently high amount of medical care. The efficiency losses from moral hazard are greater the more elastic the demand for medical care.

What determines the elasticity of demand for health care? Demand tends to be more elastic for goods that are not considered necessities and goods with substitutes. Different types of medical care are likely to differ in their elasticity; one might have more elastic demand for the doctor with a sore throat versus a heart attack, for instance. With some sore throats, one can rest, use over-the-counter remedies, and eat chicken soup without significantly worse outcomes than one would receive from a doctor's visit.

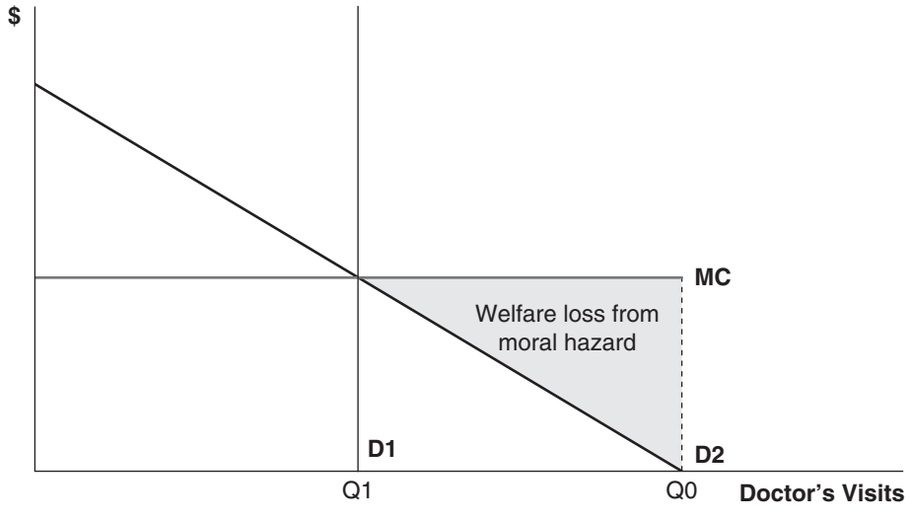


**Figure 70.2** The Welfare Economics of Moral Hazard

Even if it is a strep throat, which requires a prescription for antibiotics, consumers with no insurance may wait for a bit to determine whether the sore throat will get better on its own. Consumers with insurance are more likely to go to the doctor sooner, and in some of those cases, the doctor's care is unnecessary because the illness would have resolved by itself. Because heart attacks are more life threatening and there are not many true substitutes to medical care, demand for cardiac care is likely to be more inelastic. The RAND health insurance experiment (discussed below) provides some evidence on the elasticity of demand for different types of health care services.

The overconsumption from moral hazard is rational for the individual, who receives positive marginal benefits from the additional care. The welfare problem is that his or her marginal benefits are less than the true resource cost of production. From a society's view, this behavior is undesirable; the resources used to produce those doctor visits could be going to producing other goods with higher marginal benefits in other markets. The amount of the efficiency loss can be measured if the marginal benefits of care are known. The deadweight loss of moral hazard is shown by the area of the shaded triangle in Figure 70.2. The welfare losses are not equal to the full resource costs of production because we still do count the marginal benefits accruing to the individual.

### *Market Solution: Benefit Design*

The problem of moral hazard can be minimized by patient cost-sharing mechanisms. If designed well, the addition of deductibles and coinsurance can reduce the amount of welfare loss produced by insurance by returning some of the costs back to the patient. The introduction and encouragement of consumer-directed health care is an example of a market solution intended to combat moral hazard. These plans are typically a very high deductible catastrophic insurance plan. These insurance policies are then priced more cheaply than more generous coverage. The deductibles in these plans force the patient to face the full marginal cost of the first few thousand dollars of care. Presumably, consumers will then be less likely to go to the doctor for minor problems like colds and to think about less expensive options for other types of care.

The success of patient cost sharing at reducing use of health care presupposes that the consumers know the true marginal benefits of the health care and that they fully direct their health care expenses. The problem with these assumptions is that there is asymmetric information in the health care markets as well. Physicians and other health care professionals have considerably more technical knowledge than patients, and patients therefore rely on them to be good agents when prescribing care. In addition, can the average patient distinguish between the medical problems that require medical intervention from those that will resolve on their own? The heart attack example is illustrative; perhaps the symptoms experienced by the patient are nothing more than heartburn, but without medical personnel to assess the patient, the consumer really does not know. To the extent that physicians direct the consumption of medical care, this limits the responsiveness of consumers to price. Patients may not be presented with a range of options for a particular health problem or may not possess the information to adequately assess the different costs and benefits of different options.

Cost containment mechanisms such as deductibles and coinsurance are common in the United States but rarer and/or more modest in other countries. While there is evidence that co-payments do reduce the quantity of care demanded, in a global sense, they seem to be fairly ineffective at reducing costs for the entire health care sector. Other countries, which rely more on cost containment strategies directed at the supply of health care, have lower per capita health care costs than the United States.

### *Income Taxes and Moral Hazard*

The problem of moral hazard may be exacerbated by the differential treatment of health insurance "purchased" at work. In essence, total compensation received by workers in efficient markets is equal to the marginal revenue product that worker brings to the firm (his or her productivity). Workers could receive compensation in the form of wages or other benefits, including health insurance, pensions, and so on. To receive health insurance, workers must exchange wages for that benefit. The worker pays for the insurance directly through the employee share of the premium and indirectly in the form of lower raises and reduced wages.

The growth in employer-provided group health insurance in the United States has been fueled in part by the differential treatment of health insurance compensation versus wage and salary compensation. We are taxed on the latter but not the former at the federal and some state levels. This disparate handling in the tax code is rooted in the wage and price controls imposed during World War II. Having a large military mobilization increased both the domestic demand for goods and the domestic labor demand. Policy makers recognized that both factors would likely lead to considerable inflation. In the wartime environment, the government took a much more active role in regulating the economy, and one policy was the wage freeze. However, firms still required some means of trying to attract more workers in the tight labor market and began to offer additional, nonwage, fringe benefits. These were allowed and not considered part of the wage/salary freeze. This circumstance led to some questions about how nonwage compensation would be taxed. In 1954, the Internal Revenue Service (IRS) issued a judgment that these parts of compensation were not part of taxable income.

As a result, individuals who receive insurance at work pay less than a dollar for a dollar's worth of coverage. In effect, the treatment acts like a tax deduction; it reduces taxable income by the value of the insurance. In addition, it reduces payroll taxes. Policies such as this one are

termed *tax expenditures;* they represent foregone tax revenues for the government. How much the tax bill is reduced for the family depends on which tax bracket the household falls into. The value of a deduction is equal to the premium multiplied by the highest marginal tax rate the household faces.

Because of the artificial lowering of the price of health insurance at work, workers may demand more of their compensation in the form of health insurance than is optimal. Insurance policies traditionally covered rare, catastrophic events; now they often cover routine, smaller ticket items. In essence, health insurance now takes on an additional role beyond that of taking away uncertain risks; it becomes the predominant form of financing for health care.

### *Welfare and Equity Implications*

If no market failures existed in health insurance, the special tax treatment would artificially reduce the price of health insurance and cause inefficiently high levels of health insurance coverage. However, asymmetric information does exist. Increasing the quantity of health insurance in this context may imply that the tax treatment improves efficiency. Second, equity considerations are also considered to be relevant. Society may view access to health care as being a special good that should not be allocated by the market. Many believe that low income should not prevent access to health care. If the tax treatment of health insurance improves access to health care for disadvantaged groups, society may value that outcome even if some efficiency costs exist in increasing the equity of outcomes (Okun, 1975). The tax treatment, however, does not appear to be accomplishing the goal of increasing equity of coverage.

### *Efficiency Impacts*

Critics argue that the favorable tax treatment of employer-provided health insurance has led to an inefficiently high level of coverage. Health insurance has become less health insurance and more health care financing, with benefits including small, regular expenses such as checkups, in addition to the coverage of big-ticket, uncertain expenses. More generous benefits lead in turn to more potential for moral hazard, as patients become increasingly insulated from the full marginal cost of care. As discussed above, the amount of inefficiency caused by this moral hazard depends on how elastic the demand for medical care is.

What is the evidence on elasticity of demand for health care? Estimating the elasticity of demand for health care by varying the levels of co-payments patients pay is subject to selection bias since, as we have described above, individuals choose their level of coverage based in part on their health status. In other words, consumers who are relatively unhealthy are more likely to opt for fuller insurance coverage with lower patient out-of-pocket expenses. The RAND health insurance study provides the cleanest estimates of the elasticity of demand for health care. In this study, individuals were randomized to different levels of health insurance coverage, and their medical care use was tracked, so their level of insurance was uncorrelated with their health status. For most services, the estimated demand elasticities were quite inelastic (Newhouse & the Insurance Experiment Group, 1993). Inelastic demand for medical care then implies relatively small losses from moral hazard.

### *Equity Considerations*

How the tax treatment affects access to disadvantaged groups (the poor or the sick, for instance) is another societal concern. The tax benefits are highly inequitably distributed and make the relative access to private health insurance worse. The reasons for the inequity are twofold. Workers must first have an offer of insurance from their employer to take advantage of the tax subsidy. In addition, the size of the tax subsidy depends on the marginal tax rate facing the household receiving the insurance. Because the U.S. income tax structure is progressive, higher income households receive a higher tax subsidy for the same policy relative to a lower income household. Sheils and Haught (2004) estimate that the bulk (71.5%) of the benefits of this tax expenditure goes to families with incomes of $50,000 or more, relative to the rest of the population. Looking at the data in a different way, the average benefit for families with incomes of $100,000 or more is $2,780, compared to a mere $102 on average for families with less than $10,000 in annual income.

A horizontal inequity is also created by the policy. Otherwise similar individuals are treated differently, depending on where they access health insurance. While self-employed individuals are also now able to deduct their insurance, the treatment for individuals purchasing insurance in the nongroup market is quite different. Individuals who have medical expenses exceeding 7.5% of their adjusted gross income may deduct these expenses from their taxes; insurance premiums may count as part of these medical expenses. In other words, individuals must have substantial expenses before they are able to take advantage of this deduction. In addition, the employer-provided health insurance exemption reduces individuals' payroll tax (FICA) contributions, while the deduction for individual purchases of health insurance does not, reducing the relative value of the latter.

## Public Insurance in the United States

Now that we have a better sense of the workings of the health insurance market, we turn our attention to the public insurance policies in the United States. The U.S. system of financing is a mix of private and public coverage. While there are a variety of government-run and funded health programs, including coverage for military personnel and veterans, this chapter will focus on the three main programs: Medicare, Medicaid, and SCHIP.

### Medicare

Medicare, established in 1965, is the public coverage for individuals age 65 and older, individuals younger than age 65 with disabilities, and anyone with end-stage renal failure. In essence, for most recipients, it is a single-payer-type plan. The "traditional" Medicare policy is composed of two parts: Part A, essentially catastrophic coverage for inpatient hospital care and some temporary, limited coverage for hospice and long-term care coverage. These benefits are not subject to a premium. The second part, Part B, covers some outpatient and doctor's services. Beneficiaries contribute premiums for this coverage. Other funding is from payroll taxes and general tax revenues.

Individuals can also opt out of traditional Medicare coverage and choose Part C, aka "Medicare Advantage" plan, which could be an HMO, PPO, or a private fee-for-service plan. The government pays the premium for the individual, and the private insurance plan takes over the insurance role. With the passage of the Medicare Modernization Act of 2003, prescription drug coverage was added in 2006. Part D includes some premiums for individuals. In addition to this basic level of coverage, elderly individuals may purchase supplemental, or gap, insurance to cover expenses that are not covered by Medicare. Retiree health insurance plans often fill this role, as does Medicaid coverage for the elderly poor. Individuals who receive both Medicare and Medicaid coverage are called dual eligibles.

### Medicaid and State Child Health Insurance Program

Medicaid and SCHIP are means-tested programs to provide insurance for the very poor. They are jointly state and federally funded and administered insurance for the poor. Eligibility for Medicaid varies by state, with the federal government establishing minimum criteria. As a means-tested program, individuals must have incomes below a specific cutoff based on the federal poverty line and hold limited assets to qualify for the benefits. The income eligibility cutoffs also vary by characteristics of the individual, with higher thresholds for infants, children, and pregnant women, for example. Individuals with very high medical expenses may qualify under the medically needy category; Medicare dual eligibles often qualify through this route.

## The Uninsured

Not everyone in the United States has access to coverage. Given the market failures and the structure of public coverage, it is possible to predict characteristics that will be associated with lack of insurance. Individuals with preexisting conditions who do not have access to the group market will be particularly vulnerable. The elderly receive coverage through Medicare, so the uninsured will be found among the nonelderly. Means-tested programs cover the very poor, and to receive health insurance compensation through an employer, the worker's marginal revenue product must be high enough to pay for the insurance in lower wages.

The uninsured are by and large in families with one or more full-time workers and in families that are under or near the poverty line. They earn too much income to qualify for Medicaid/SCHIP or do not take up Medicaid, but they do not have enough income to purchase health insurance on their own. Since the benefits of risk pooling increase with the size of the group, uninsured workers are more likely to be employed at smaller firms, in occupations and industries with considerable job turnover, and in low-wage occupations.

## The Evolution of HIV/AIDS and Insurance Coverage

On June 5, 1981, the Centers for Disease Control (CDC) reported a small cluster of deaths in Los Angeles from a form of pneumonia that was rarely fatal. After long investigations, the underlying cause was determined. Acquired immune deficiency syndrome (AIDS) was identified, an infectious disease caused by a retrovirus named human immunodeficiency virus (HIV). In the early years of the epidemic, HIV was a virtual death sentence, and until a blood test was approved by the Food and Drug Administration (FDA) in 1985, most cases were not identified until late in the disease stage, when the immune system had already been so compromised that opportunistic infections had already attacked the system. The disease has a long latency period, on average 6 to 10 years, in which the individual is infected with the virus before developing full-blown AIDS. During the latency period, the individual may show few or no debilitating symptoms. An individual develops AIDS when his or her CD4 cell count (an important component of the immune system) drops below 200 and/or begins to contract opportunistic infections.

In the beginning of the epidemic, treatment of the opportunistic infections was the only course of action available until the introduction of azidothymidine (AZT), which inhibits the replication of the virus itself. However, because of HIV's quick replication rate and high rate of mutations, AZT alone only temporarily extended the life expectancy of those infected before the virus developed resistance to the pharmaceutical. For many during this period, life expectancy was extended only a few months. It was not until the development of highly active antiretroviral therapy (HAART) in 1995 that medicine was able to significantly extend the life expectancy of HIV-positive individuals and the death rate from AIDS began to fall. HAART is characterized by treatment with more than one drug; each drug targets

a different aspect of the viral infection and replication process. Success of the regimen depends on the strict adherence to the treatment; treatment cannot be stopped once started, or the virus may gain resistance, and the disease progression will resume.

The evolution of treatment for HIV/AIDS and the characteristics of the disease itself provide an important policy problem within the United States for the study of health insurance. As we have examined above, insurance "for insurance's sake" should be focused on catastrophic (high-cost), unexpected health problems. AIDS fit this model quite well early on in the epidemic since it was an acute problem, although the outcome was often death and not recovery. It was not until the development of effective medical therapies that HIV/AIDS changed to a chronic condition with important implications for the financing of health care for HIV-positive individuals.

Once started, to be effective, the drugs must be taken daily for the rest of the patient's life, to prevent the development of drug resistance. However, once an individual has become diagnosed with HIV, the level of uncertainty about the amount and costs of medical care required by the individual drops. One faces *certain* medical expenditures for many, many years. The HAART treatments alone are quite expensive, averaging $12,000 to $24,000 per year, not counting other treatments for opportunistic infections and toxicity-related side effects of the drugs themselves.

Individuals who have tested positive for HIV experience enormous difficulties in obtaining private coverage in the individual market. The presence of the virus is a preexisting condition with known, large current and future expenses. Pollitz, Sorian, and Thomas (2001) surveyed nongroup market insurers about premiums and policy restrictions for a set of hypothetical patients. In this study, even individuals who only had hay fever as a preexisting condition faced higher premiums and fewer offers of "clean" coverage (coverage without additional restrictions on the benefits); the hypothetical HIV-positive consumer received no offers of coverage at any price in any of the markets. To receive access to private group coverage, an individual must be able to work, find a job/employer that offers coverage, and be able to afford the employee share of the premium. If HIV/AIDS results in a disability for the individual, employment will become difficult or impossible.

The typical routes to public financing for HIV/AIDS care are through either Medicare under the low-income, disabled eligibility category or Medicaid as a low-income welfare, medically needy, or as a dual-eligible Medicare disability recipient. To qualify as a person with disabilities, one must have an employment history and wait for 2 years after the development of the disabling condition to become eligible. Once on the HAART therapy, HIV/AIDS can become less disabling, which then may disqualify the person from eligibility for the public coverage. If public coverage is lost, the patient may stop taking the drugs, become disabled again, and can then regain coverage. From a public health perspective, this on-again/off-again financing of care could contribute to the rise of drug resistance in the virus, a significant negative externality. See Laurence (2001) for further analysis of the policy debate.

Given the market solutions to adverse selection and the strong tie of private insurance to employment, one would expect to find a lower rate of private insurance coverage among the population of HIV-positive individuals. This trend away from private coverage toward public coverage should increase over time, as medical technology has transformed HIV/AIDS into a chronic, albeit often debilitating, disease model. Data in Goldman et al.'s (2003) analysis confirms this picture, as half of all HIV-positive individuals in the study are insured by Medicaid, Medicare, or both. The importance of the public sector as a safety net for persons living with HIV has disadvantages. During recessionary time periods, states face pressure to cut back on their outlays for Medicaid, which can hurt access to care for this population. Ghosh, Sood, and Leibowitz (2007) find that one state strategy for cutting costs, decreasing the income threshold to qualify for Medicaid, increases the rate of uninsurance and decreases the use of HAART.

Stability of coverage is another issue of great importance for HIV-positive individuals. To what extent do HIV-positive individuals move from being insured to losing coverage? How often do they switch from private to public coverage or vice versa? Do these transitions affect the quality of medical care received by this population? Fleishman (1998) provides some evidence from the 1990s; most individuals who changed coverage status were those who went from no insurance to public insurance; individuals who had developed AIDS were more likely to make such a switch. Relatively few individuals in this sample lost private coverage, perhaps reflecting the increase in the effectiveness of HAART therapy at keeping individuals from developing disabling symptoms. More recent evidence in Kelaher and Jessop (2006) in New York City also shows the same pattern. However, some individuals do lose coverage, which has important implications for their ability to afford therapy.

## Conclusion

Health insurance markets do not fit the standard market model of perfect competition. They exist because of the information problem of uncertainty, in combination with our risk aversion. Once created, they face the market failure of asymmetric information. Being insured may alter the consumption of health care in inefficient ways through moral hazard. All of these problems have real consequences for who gets covered and how they are covered. When analyzing these markets, society considers not only the efficiency of the market, but because insurance

finances health care, equity concerns are also prominent in the debate on the appropriate policy response.

## References and Further Readings


Akerlof, G. A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics, 84,* 488–500.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review, 53,* 941–973.

Centers for Disease Control. (1981). Pneumocystis pneumonia—Los Angeles. *Morbidity and Mortality Weekly Reports, 30,* 250–252.

Cutler, D. M., & Gruber, J. (1996). Does public insurance crowd out private insurance? *Quarterly Journal of Economics, 111,* 391–430.

Fleishman, J. A. (1998). Transitions in insurance and employment among people with HIV infection. *Inquiry, 35,* 36–48.

Ghosh, A., Sood, N., & Leibowitz, A. (2007). The effect of state cost containment strategies on the insurance status and use of antiretroviral therapy (HAART) for HIV infected people. *Forum for Health Economics & Policy, 10*(2), Article 3. Retrieved February 24, 2009, from http://www.bepress.com/fhep

Goldman, D. P., Leibowitz, A. A., Joyce, G. F., Fleishman, J. A., Bozzette, S. A., Duan, N., et al. (2003). Insurance status of HIV-infected adults in the post-HAART era: Evidence from the United States. *Applied Health Economics and Health Policy, 2*(2), 85–91.

Kelaher, M., & Jessop, D. J. (2006). The impact of loss of Medicaid on health service utilization among persons with HIV/AIDS in New York City. *Health Policy, 76,* 80–92.

Laurence, L. (2001, March 13). Special report: Medicaid's catch-22. *Kaiser Daily HIV/AIDS Report.* Retrieved February 24, 2009, from http://www.kaisernetwork.org/daily_reports/rep_index.cfm?DR_ID=3349

Medicare Payment Advisory Commission (MedPAC). (2008, December 5). *Medicare advantage program.* Paper presented at the MedPAC Public Meeting, Washington, DC.

Newhouse, J. P., & the Insurance Experiment Group. (1993). *Free for all? Lessons from the RAND health insurance experiment.* Cambridge, MA: Harvard University Press.

Okun, A. (1975). *Equality and efficiency: The big tradeoff.* Washington, DC: Brookings Institution.

Pollitz, K., Sorian, R., & Thomas, K. (2001). *How accessible is individual health insurance for consumers in less-than-perfect health?* Retrieved February 24, 2009, from http://www.kff.org/insurance/upload/How-Accessible-is-Individual-Health-Insurance-for-Consumer-in-Less-Than-Perfect-Health-Report.pdf

Sheils, J., & Haught, R. (2004). The cost of tax-exempt health benefits in 2004. *Health Affairs.* Available at http://content.healthaffairs.org/index.dtl

Summers, L. H. (1989). Some simple economics of mandated benefits. *American Economic Review: Papers and Proceedings, 79,* 177–183.

Woolhandler, S., Campbell, T., & Himmelstein, D. U. (2003). Costs of health care administration in the United States and Canada. *New England Journal of Medicine, 349,* 768–775.

Woolhandler, S., & Himmelstein, D. U. (2002). Paying for national health insurance–and not getting it. *Health Affairs, 21*(4), 88–98.

# 71

# Economics of Information

Viktar Fedaseyeu
*Boston College*

As Francis Bacon once noted, knowledge is power. The need to acquire and process information is a tremendously important part of human interactions. Economic interactions are no exception. Investors buying stocks or other securities need to analyze the quality and reliability of information supplied to them. Employers must evaluate qualifications of job applicants prior to starting a working relationship with them. Insurance companies must set premiums based on the perceived risks of the insured. Examples of this sort abound in any field of economics.

The purpose of this chapter is to show that the structure of the information environment can have significant consequences for how the markets and the whole economy operate. Standard theories of competitive equilibrium usually avoid these issues by assuming that all agents have costless access to the same information. The recognition that this is a very limiting assumption is what has fueled research in information economics.

Information, unlike other goods that interest economists, has several unique characteristics:

- Information is a public good: Learning does not prevent others from learning.
- Information cannot be unlearned; that is, the decision to acquire information is irreversible once it has been acquired.
- Information is asymmetric: Different people know different things.

It is conceptually easy to understand the public-good nature of information. The other two characteristics require further clarification. The fact that information cannot be unlearned complicates the process of information production. Unlike other goods, information cannot be sampled or returned for a refund: Once the decision has been made to learn something, it is virtually impossible to reverse it after obtaining information. It is of course possible to forget information. Forgetting, however, will not reimburse one for the efforts expended to obtain information.

Information asymmetry, the third property listed above, will be the primary focus of this chapter. The idea that people have different knowledge seems almost obvious. Its implications for market equilibrium, however, had not been recognized for a long time. It was usually assumed that in competitive markets, knowledge heterogeneity will be small and will not affect the equilibrium in any significant way. Contrary to this common belief, Akerlof and Yellen (1985) showed that even small deviations can have significant effects.

George Stigler is often regarded as the father of economics of information for it was he who focused economists' attention on the role of information in economic decision making. Stigler (1961) analyzed the dispersion of prices and noted that while it is often convenient to assume that all agents are endowed with perfect information about preferences and technologies, the very existence of price dispersion contradicts this simple view. He explained the persistence of such dispersion by the presence of search costs: Obtaining information requires time, which is a valuable resource. He then suggested that in equilibrium, the marginal benefit of the search for information must equal its marginal cost.

While Stigler's (1961) paper was a breakthrough that gave birth to a new field in economics, it treated information in very much the same way that the economists were treating other goods. Information, however, is different in many respects. It is these differences, and information asymmetry in particular, that will be the focus of our attention.

## Information Asymmetry

### The Problem of Asymmetry

Information is valuable, and its value is often determined by the number of people who know it. The fact that some people know more than others is referred to as information asymmetry. The examples of asymmetry are numerous: A job applicant knows more about his or her abilities and work ethics than the potential employer, an entrepreneur seeking a loan from the bank knows more about the viability of his or her business than the loan officer, and a person buying health insurance knows more about his or her health than the insurance company. Asymmetry is a departure from the model of perfect markets, and it creates challenges and problems.

While the problems of information asymmetry had been realized by early economists, Akerlof (1970) was the first to rigorously analyze the problem of private information. He showed that information asymmetry is crucial to understanding how markets operate.

Consider the market for used cars, which is the example discussed in Akerlof (1970). Assume that a car can be of either good or bad quality with equal probability (½)—that is, there are as many good cars as there are bad cars. Further assume that for the sellers, a good car is worth \$800 and a bad car is worth \$50. For potential buyers, on the other hand, a good car is worth \$1,000 and a bad car is worth \$100. What is the equilibrium clearing price in this market? As it turns out, the answer depends on who knows what.

First, consider the case when neither sellers nor buyers can differentiate between good and bad cars. The expected value to the seller of not selling the car is \$425 = ½ × \$800 + ½ × \$50. The expected purchase value to the buyer is \$550 = ½ × \$1,000 + ½ × \$100. Thus, at a price between \$425 and \$550, cars of all qualities will be traded in this market.

Now assume that sellers know the quality of their cars while buyers do not, which seems quite realistic. Notice that there is no credible way sellers can communicate their knowledge since every seller would try to claim that he or she has a good car. The \$425 to \$550 range no longer works in this case because sellers with good cars will be unwilling to sell them for less than \$800. Therefore, only bad cars will be brought to the market. The buyers understand this and will only pay \$100 or less since this is the value of bad cars to them. Notice that many welfare-enhancing trades will be forgone in the above equilibrium. A market breakdown occurs.

### Adverse Selection and Moral Hazard

The applicability of Akerlof's (1970) result is very wide. Research that followed his seminal paper identified two major types of problems that arise from information asymmetry: adverse selection and moral hazard. The problem that Akerlof originally described in his work is a special case of adverse selection. It is now often associated with insurance markets, although it is far more general.

Consider the problem of selling health insurance. Different people have different degrees of health risks, and most important, they are likely to know more about their health than insurance companies. Notice that consumers with higher risks will value insurance more and will be more likely to buy it. As a result, the proportion of less healthy people in the pool of actual insurance buyers will be larger than their fraction in the general population. It is the more risky consumers who self-select themselves into buying the product—hence the term *adverse selection.* It is, of course, the opposite of what insurance companies would want.

Moral hazard refers to the inappropriate behavior of agents whose actions cannot be perfectly monitored. A person who has car insurance may be less vigilant about the car than he or she would have been without insurance. In this case, the insurance company has less information about future actions of the person than the person himself or herself. The principal-agent problem is a special case of moral hazard: As long as the principal cannot monitor the agent perfectly, the latter has an incentive to engage in behaviors inconsistent with the principal's interest. A simple albeit extreme example will illustrate this. Imagine a principal who hires an agent to manage retirement savings and agrees to pay the latter a fixed salary no matter what. It is obvious that the agent's optimal behavior would be to do nothing at all and simply collect the salary without effort.

Government welfare policies may sometimes create moral hazard problems. Unemployment benefits are one example: As long as a person is qualified to receive them, he or she may be less willing to search for a job. Another example is government bailouts. An expectation that the government will help some companies in bad times may create an implicit guarantee. As a result, such companies may undertake projects that they would have considered too risky had there been no implicit guarantee. While in many instances, letting a business entity fail may have catastrophic consequences, policy makers and students of economics should be aware of moral hazard implications and take them into account in their decision making.

Akerlof's (1970) pioneering work outlined the problems that can arise in markets with asymmetric information. It did not, however, look at how the markets can address these problems. Later research by Michael Spence and Josef Stiglitz suggested two elegant approaches: signaling and screening.

### Signaling

Henceforth in our discussion, we will refer to agents with private information as the informed party, while their counterparty will be called the uninformed party.

Spence (1973) introduced the concept of signaling. While his original paper discussed a job market signaling model, its general idea can be applied in a wide variety of settings. The premise of Spence's reasoning is that informed agents can behave in a way that sends a "signal" to the uninformed party. This behavior can be used to reveal information about product quality, a person's qualification for the job, and so on. The signal, however, must be credible. Credibility is usually associated with costs for the informed party: If there are no costs involved, everyone would try to send the signal with the most favorable information, and the Akerlof-type market breakdown is likely to occur. The costs must be high enough to prevent some types of agents from sending the signal and low enough to make signaling worthwhile for other types. If this condition is met, signaling will separate agents based on their types and thus reveal information.

In Spence's (1973) original paper, schooling is used by potential employees to signal the level of their ability. Assume that there are two types of workers: those who are highly qualified for the job and those who are less qualified. The employer is unable to differentiate them a priori. The workers, however, are aware of their ability level. In a competitive labor market with complete information, every worker's wage will be determined by his or her marginal productivity. In the incomplete information case, on the other hand, the employer is unable to perfectly determine the worker's marginal product prior to hiring and must take into account the probability that the worker is unskilled.

The critical assumption in Spence (1973) is that marginal costs (monetary and mental) of schooling are greater for workers with less ability. As a result, the level of schooling is positively correlated with ability. What makes the signal credible is the fact that beyond a certain level of education, less able workers will find it unjustified to continue acquiring more schooling: Their costs will exceed potential benefits from higher wages. It is only the more able applicants who will find it worthwhile to pursue further education. Thus, additional investment in education will differentiate applicants based on their ability.

The cost of schooling in Spence's (1973) model is dissipative: Education does not bring any benefits beyond signaling the worker type. And since highly qualified workers have to privately bear the costs of education, total welfare declines relative to the full-information case. It is likely that other benefits of education may affect marginal productivity directly. In particular, even less skilled applicants can acquire necessary qualifications through education. These benefits are clearly important, but they may not solve the asymmetry problem since more able workers may still remain more qualified compared to the less able applicant with the same level of schooling.

## Screening

Although conceptually similar to signaling, screening is a slightly different device used by the uninformed party to extract information. The uninformed agent may offer a menu of options to the informed party, whose choices out of that menu reveal information that the uninformed agent is seeking. Notice that this menu must be somewhat restrictive since letting the informed party choose whatever actions they prefer will not alleviate the problem.

Rothschild and Stiglitz (1976) developed a classical screening model for insurance markets. People who buy insurance usually know more about the risks they are facing. They will, therefore, self-select into buying insurance for the kinds of losses that are likely to occur to them. Insurance companies must account for this in the premiums they charge. In particular, premiums must be higher than they would have been had every person bought insurance regardless of his or her risk level because the proportion of high-risk individuals in the general population will be lower than among those who actually buy insurance. These higher premiums, however, will attract even fewer low-risk individuals and will only exacerbate the self-selection problem. Rothschild and Stiglitz show that competitive insurance markets can only be in equilibrium if insurance companies compete on both price and quantity—that is, they offer contracts that specify both price and quantity of insurance that can be purchased for that price. This is different from the usual competitive market equilibrium where the seller offers a price and the buyer determines the quantities he or she is willing to obtain.

Do real-life insurance markets behave in a way that is consistent with above conclusions? They do. In particular, insurance companies offer various quantities of coverage by using deductibles and caps. A deductible is the amount that the insured person must cover himself or herself, while any damage beyond that amount is paid by the insurer. A cap is the maximum possible amount of insurance that the company agrees to pay. Individuals facing higher risks will prefer lower deductibles and higher caps, while low-risk people will be willing to buy insurance with higher deductibles and lower caps. As one might expect, the magnitude of the deductible and the cap a person is willing to accept affect premiums very significantly.

Rothschild and Stiglitz (1976) show that under certain circumstances, insurance markets can have no equilibrium. If feasible equilibrium exists, however, it can only be separating. In a pooling equilibrium, people of all risk types buy the same contract, and screening is unnecessary. In a separating equilibrium, on the other hand, every type buys the contract geared specifically to that type's risk profile. Separating equilibria are usually welfare inferior to pooling equilibria because of screening costs. Rothschild and Stiglitz show that low-risk people are left underinsured compared to the full-information case. This fact prompted policy makers to advocate mandatory insurance to mitigate the adverse selection problem. The reasoning goes as follows. If everyone is required to purchase insurance, people will no longer have the option to self-select based on their risk characteristics. Insurance companies will know that

people buying the contracts were required to do so regardless of their risk characteristics and may be willing to lower the rates to reflect the expected average level of risk. This average level of risk will be lower than without mandatory insurance since people of different risk types are required to participate. The effectiveness and fairness of this approach is an important part of current policy debates.

It should also be clear that insurance markets might have other problems that may diminish the effectiveness of this policy. In particular, Rothschild and Stiglitz (1976) forego a detailed discussion of the incentives of insurance companies themselves, mainly because of the lack of theoretical guidance in that respect.

## Applications: Financial Markets

### Credit Rationing and Screening With Interest Rates

One of the puzzling features of economic reality is the fact that credit is rationed. In simple terms, credit rationing is the situation when credit markets do not clear (i.e., the demand for funds exceeds their supply at the prevailing interest rate). As Stiglitz and Weiss (1981) note, when the price mechanism works well, rationing should not exist: If demand for credit were to exceed supply, prices (i.e., interest rates) would rise and lead to an appropriate increase in the supply of funds. In Stiglitz and Weiss, it is information asymmetry that may be responsible for credit rationing.

Interest rates may affect the riskiness of the pool of potential borrowers, in which case the problems of adverse selection and moral hazard arise. The observation that high interest rates may drive away the safest borrowers dates back to Adam Smith. Those who are willing to accept higher interest rates may be, on average, of higher risk to the bank since they may be willing to borrow at higher rates precisely because they perceive their repayment probability as being low. In addition, higher interest rates may lead to moral hazard and change the behavior of the borrowers: Stiglitz and Weiss (1981) show that higher interest rates lead firms to undertake projects with lower probabilities of success but higher payoffs conditional on success. As a result, the bank's expected profits may decline after a certain cutoff interest rate. It is conceivable that at this cutoff rate, the demand will still exceed supply, but banks will be unwilling to lend due to the moral hazard and adverse selection problems.

### Information Asymmetry Problems in Corporate Finance

#### *Signaling With Capital Structure*

Capital structure is a focal problem in corporate finance. Academics have for a long time been concerned with why firms choose a particular mix of debt and equity financing. The very fact that firms opt to manage their capital structure is puzzling from the standard viewpoint of competitive markets. Modigliani and Miller (1958) showed that in a frictionless world, the choice of debt over equity should not matter. Empirical studies, on the other hand, have documented several stylized facts about leverage, and these facts required explanation. Several theories and a strand of empirical papers testing those theories emerged (for an incomplete list, see Graham, 2000; Rajan & Zingales, 1995; Titman & Wessels, 1988).

Myers and Majluf (1984) discuss corporate structure decisions in a setting of incomplete information: The firm's insiders know more about the firm's assets than outside investors. Assume that the firm has a profitable investment opportunity and must raise outside financing in order to pursue it. Also assume that the management acts in the interest of existing shareholders. Due to information asymmetry, the market will at times undervalue the firm relative to insiders' private information. If new equity is issued, existing shareholders will be diluted since they will have to share profits with new shareholders. The extent of this dilution will be large if the insiders have very favorable information about the firm relative to the market's assessment. Sometimes this dilution will be sufficiently large to result in a wealth loss for existing shareholders.

A numerical example will make this argument clear. Imagine that the market values Firm A at $1,100,000 while insiders observe the true value of $2,500,000. Firm B has the same market valuation, but insiders value it at $1,500,000. Assume that both firms have access to identical projects that cost $320,000 and will create a net value of $180,000 next period (the gross value of the project is therefore $500,000). Also assume that information asymmetry will be resolved next period regardless of whether the project is implemented. If both firms implement the project, Firm A will be worth $3,000,000, and Firm B will be worth $2,000,000. The a priori market valuation, however, is different because of asymmetry: It is $1,600,000 for both firms. Thus, outside investors will be willing to provide $320,000 in exchange for a 20% stake in either company ($320,000/$1,600,000 = 0.20). Notice, however, that from the insiders' point of view, if the project is implemented, a 20% stake is worth $600,000 and $400,000 in Firm A and Firm B, respectively. Insiders are diluted in both cases, but Firm A insiders are clearly worse off.

It is, however, incorrect to assume that offering a stake will result in a wealth loss in both cases. Consider Firm B: Without the project, it will be worth $1,500,000. With outside financing, current shareholders will own 80% of a bigger $2,000,000 firm, for a total of $1,600,000. This is clearly greater than $1,500,000 they would have should they decide to forgo the project and raise no equity. Thus, even though they are forced to sell equity below its true value, the greater future value of the firm will compensate them for that. This is not the case for Firm A, however. After issuing equity and implementing the project, its current shareholders will be left with 80% of a $3,000,000 company, which translates into $2,400,000. This is less

than the $2,500,000 standalone value of current assets. Thus, Firm A will refrain from issuing equity and forego the investment.

In essence, not issuing stock is a good signal to the market since it implies that insiders have very favorable information relative to the market's assessment. To obtain a more precise intuition for this result, we provide a semi-formal discussion of a simplified example in the spirit of Myers and Majluf (1984).[1]

The assumptions are:

1. Agents are risk neutral.
2. Insiders (managers) have private information about the true value, π, of the firm's existing assets. π can have the value of either *H* or *L:*

$$\pi \in \{H, L\},\ H > L > 0.$$

3. The firm obtains access to a value-enhancing project, which will yield cash flow *C* at Time 1. The firm needs external financing *I* to undertake the project. Assume that equity is the only available option to raise funds. In principle, we could relax this assumption and decrease the amount of required investment. In other words, one can interpret *I* as the amount of required equity financing in excess of other sources available. If the firm does not issue new equity, the new project is not implemented since other sources are assumed to have been depleted.
4. Since the project has a positive net value, its cash flows must exceed cash flows obtained from comparable passive investments, or equivalently, $C > I(1 + r)$, where *r* is the return on comparable investment opportunities. For simplicity, we let *r* equal the risk-free rate.

Agents make their decisions at Date 0 after observing their private signals and the firm's actions. At Date 1, asymmetry is resolved, and cash flows are distributed. We should be careful not to attach any calendar meaning to Date 0 and Date 1 since it may take weeks or even months for asymmetry to be resolved. At Date 0:

- Insiders observe the realization of π, and outsiders observe only a probabilistic distribution $P(\pi = L) = p$ or, equivalently, $P(\pi = H) = 1 - p$.
- Insiders now offer fraction *s*, $(0 \le s \le 1)$, of the firm's shares to outside investors, in return for the amount *I.*

After observing *s,* outside investors may accept or reject the offer.

If investors reject the offer:

- Investors' payoff is I(1 + r).
- Insiders' payoff is p.

If investors accept the offer:

- Investors' payoff is $s(\pi + C)$.
- Insiders' payoff is $(1 - s)(\pi + C)$.

There are two types of equilibria in this game: the pooling and the separating equilibria. We will discuss them below. While we do not formally define the equilibrium concept used here, it may be useful to know that we are describing pure-strategy perfect Bayesian equilibria. Informally, this concept requires that actions of each agent are an optimal response to *equilibrium* actions of the other players and that in equilibrium, actions are consistent with agents' beliefs (i.e., if everyone believes a firm to be of high quality, it should take actions that a high-quality firm is supposed to take rather than mimic some other firm type).

In a pooling equilibrium, investors' probability assessment of the firm being type *L* remains unchanged after observing

$$P(\pi = L|s) = P(\pi = L) = p.$$

Equivalently,

$$P(\pi = H|s) = P(\pi = H) = 1 - p.$$

In this case, both types of firms offer *s,* which is accepted by the investors.

What are the conditions under which this equilibrium can exist? Intuitively, all parties must be satisfied with the equilibrium outcomes and beliefs and be unwilling to deviate. More formally, investors' participation constraint (they must be no worse buying equity than investing in the risk-free asset) is as follows:

$$s[pL + (1 - p)H + C] \ge I(1 + r). \tag{1}$$

Insiders' optimality requirement (they expect to be no worse off after dilution than before; notice that while their holdings are diluted after the issuance, the firm is worth more because of the profitable investment):

$$(1 - s)(\pi + C) \ge \pi \Rightarrow s \le \frac{C}{\pi + C} \text{ for } \pi \in \{H, L\}$$

This translates into

$$s \le \frac{C}{L + C}, \tag{2a}$$

$$s \le \frac{C}{H + C}. \tag{2b}$$

Equation 2b is the stronger condition (since $H > L$), and Equation 2a can therefore be ignored since it is automatically satisfied as long as Equation 2b holds. Combining Equations 1 and 2b yields the necessary condition for equity issuance to occur in a pooling equilibrium:

$$\frac{I(1 + r)}{pL + (1 - p)H + c} \le \frac{c}{H + c}. \tag{3}$$

It is easy to see that when *p* is small (close to zero), Equation 3 is almost always satisfied, and a pooling equilibrium

is feasible. The intuition behind this result is straightforward. Notice that in a pooling equilibrium, type *H* firms subsidize type *L* firms because every company gets the same average valuation, so that type *H* firms are undervalued while type *L* firms are overvalued. In essence, by participating in a pooling equilibrium, type *H* firms accept lower valuations and thus transfer part of their value to type *L* firms. For equity issuance to occur, the cost of subsidizing the type *L* firm must be compensated by the benefit of a higher firm value from the new project, which would otherwise have to be abandoned.

If *p* is high enough (i.e., the likelihood of encountering a type *L* firm is high), the pooling equilibrium is no longer feasible. It is obvious that if a separating equilibrium exists, it is the type *L* firm that makes stock issuance since costs of issuance are higher for type *H* firms. In a separating equilibrium, therefore, the market must correctly anticipate this and update the probability of firm type based on observed behavior: If the firm announces an issuance, investors set the probability of it being type *L* equal to 1 or, equivalently, $P(\pi = H \mid \text{Issuance announcement}) = 0$.

In such equilibrium, the type *L* firm issues equity with

$$s' = \frac{I(1+r)}{L+C}$$

(i.e., Equation 1 is satisfied as an equality, with $p = 1$). Investors accept this offer since they are no worse off than investing in the riskless asset. Type *H* firms do not issue stock and do not undertake the project since issuing stock and diluting current shareholders will reduce their wealth beyond what could be compensated by profits from the new investment.

Notice that in this equilibrium, only the type *L* firm implements the project. As a result, social welfare declines since some firms have to forego profitable investment opportunities. To avoid such a scenario, firms may accumulate what Myers and Majluf (1984) call "financial slack"—that is, retained earnings and the capacity to issue riskless debt: If issuing new shares has dilutive effects and may prevent the firm from investing, it may use internal funds to undertake the investment when outside financing is unattractive.

The model above explains why share prices, on average, drop when new stock issuance is announced (Mikkelson & Partch [1986] document a 3% to 4% price decline in the 2-day window around the stock issuance announcement). Myers and Majluf propose what is now called the pecking order theory of capital structure: When possible, firms should use their retained earnings to finance projects; if internal funds do not suffice, firms should issue debt, while equity financing is the option of last resort.

*Dividends*

Dividends are another type of corporate financial policy that is difficult to explain in terms of frictionless markets. If a firm has profitable investment opportunities, the managers should pursue them and thus maximize the total firm value. Distributing dividends to shareholders may indicate the lack of such investment opportunities. Surprisingly, however, corporations issue dividends and obtain external financing at the same time, thus distributing funds to their shareholders and making investments at the same time. Miller and Rock (1985) address this problem by considering the information content of dividends. Dividends are a signaling device because of the commitment associated with them: Dividends must be paid at regular intervals and therefore effectively restrict free cash flows of the firm. In equilibrium, firms with relatively high cash flows can afford to pay dividends, and thus dividend announcements can be used to signal firm value.

## Market Efficiency

Stock prices communicate information about the firm's expected future performance. Determining the amount of information that a stock price contains is therefore tremendously important since it underlies capital allocation decisions in the economy. The question is, "What does the stock price really tell us?"

Fama (1965) advocated the concept of informationally efficient financial markets, meaning that stock prices reflect all information currently available. Efficiency does not assume that prices do not change over time or that they correctly reflect true intrinsic values of securities. It does assume that prices are the best guesses about intrinsic values currently available. In a later paper, Fama (1970) distinguished three different forms of the efficient market hypothesis: the weak, semi-strong, and strong market efficiency. Weak efficiency states that future stock prices cannot be predicted from past prices or returns, semi-strong efficiency exists when prices quickly reflect all available public information, and strong-form efficiency implies that both private and public information is reflected in the current stock price. Fama argues that in modern financial markets, many analysts try to extract information about stocks. It is the competition between them that leads to new information being incorporated into prices quickly: If somebody perceives that there are profitable trades to be made, those trades will be executed. Fama (1970) lists the following sufficient conditions for markets to be fully informationally efficient:

1. There are no transaction costs.
2. All available information is costlessly available to all market participants.
3. Everyone agrees on the implications of current information for the current price and the distribution of future prices.

While Fama (1970) concedes that these conditions are extreme and are likely to be violated in real markets, he states that they need to hold to a "sufficient" degree to ensure efficiency.

To make the model of efficient markets empirically testable, it must be given statistical content. Usually, market

efficiency has been associated with the random walk hypothesis, which states that successive price changes (or returns) are independently and identically distributed. Empirical evidence accumulated from the 1960s to 1990s was generally supportive of the efficient market hypothesis, at least in its semi-strong form. In particular, successive price changes were found to be almost independent, corroborating the random walk hypothesis. More recent evidence, however, is mixed. For example, low price-to-earnings stocks tend to outperform other stocks in a demonstration of what appears to be inconsistent with even the weak-form efficiency.

The primary problem with tests of market efficiency, which Fama (1991) acknowledges, is the joint hypothesis test: Efficiency can only be tested relative to a normative model of stock price behavior. In the absence of such a model, it is virtually impossible to statistically determine whether it is the model that is incorrect or the market that is inefficient.

Grossman and Stiglitz (1980) challenge the assumptions underlying the efficient markets hypothesis. They develop an equilibrium theory that explains why prices cannot at all times reflect all information available to market participants. They show that what Fama (1970) lists as conditions sufficient for efficiency are also the conditions necessary for it to exist. Below we provide a simplified version of their argument.

In a world where the search for information is costly, an investor would only expend effort to acquire it if he or she can earn compensation for doing so. Thus, the very decision to obtain information is endogenous and depends on the information content of prices. If prices always reflect all available information, searching for information elsewhere does not provide any value and cannot be justified. On the other hand, if nobody is trying to acquire information beyond prices, these prices cannot reflect all available information by definition, and thus abnormal profits can be made. As a result, what Grossman and Stiglitz (1980) call an "equilibrium degree of disequilibrium" arises, where prices transmit information available to arbitrageurs but only partially, so that the efforts to obtain private information are justified.

## Financial Intermediation

The student of economics may wonder why financial institutions play such a crucial role in modern economic systems: In the end, it is the real productive capacity of the economy that matters. What is the role of financial intermediaries that makes them so indispensable? To address this question, we need a specific definition of an intermediary. For the purposes of this discussion, it is an institution that holds one class of securities and sells securities of other types. One example is a bank taking deposits and providing loans; another would be a company holding individual mortgages and selling bonds backed by those mortgages (the so-called asset-backed securities) to outside investors.

Financial intermediation can arise because of transaction costs: It may be difficult for ultimate borrowers to deal with lenders directly because of search and evaluation costs. However, simple transaction costs have failed to explain the magnitude of financial intermediation: While clearly present, these costs just do not seem to be large enough. Leland and Pyle (1977) suggest that information asymmetry may be the primary reason financial intermediation exists.

There are classes of assets, such as mortgages and insurance, for which it is possible to obtain private information by expending effort. A loan officer evaluating a mortgage applicant may be able to obtain quite extensive documentation about his or her financial situation, for example. In the presence of economies to scale, it may be beneficial to create organizations that specialize in collecting information about particular asset types. Reselling this information to ultimate lenders, however, can be problematic. As Leland and Pyle (1977) note, two problems arise in this case:

1. Information is a public good, so the collector will be able to capture only a fraction of its value to potential buyers.
2. Selling information is related to the credibility of that information.

The second problem is arguably more severe: The Akerlof-type market breakdown can occur. If potential buyers of information cannot distinguish between good and bad information, they will be willing to pay only the price that would reflect the average quality of information. Thus, providers of high-quality information will be unwilling to participate, and the quality of information supplied will further decline.

These problems can be overcome if the information-gathering entity becomes an intermediary, holding assets on its own balance sheet. In this case, the returns to information collection will be captured in the returns to the portfolio of assets and will therefore be privatized. The asymmetry problem can be solved by signaling, and Leland and Pyle (1977) suggest a particular type of signal that can be used in this case: Insiders can hold a relatively large share of their own firm to signal its quality. Their argument is straightforward: Risk-averse insiders would like to diversify their holdings, and it is therefore suboptimal for them to hold large shares of their own company unless they have favorable information about it.

## Reputation

Reputation may be viewed as another endogenous response to asymmetric information. It arises in a repeat-business context, with the idea that repeated interactions can support equilibria impossible in a static transaction. In a dynamic game, it may be possible to create a punishment

code severe enough so that deviating from trustworthiness is unprofitable. Creating such codes has been a challenging game-theoretic problem: The difficulty arises because the punishment code must be severe enough to prevent deviations, but at the same time, it must be credible so that those who threaten to use it will actually do so (in terms of economic rationality, it is not a good idea to hurt someone else and suffer at the same time). Abreu, Pearce, and Stacchetti (1990) develop a general treatment of this problem and show how optimal punishment codes can be constructed.

The idea of reputation in a repeat-business context is intuitively appealing. Consider a bank that tries to attract deposits. If its only interest is the transaction at hand, the optimal policy is to expropriate the depositor to the largest extent possible (within legal bounds). On the other hand, if the bank is interested in inviting future transactions from its clients, it must maintain reputation of a fair dealmaker. If potential future benefits exceed the profit foregone by not expropriating clients today, the bank has incentives to heed to reputational concerns. In essence, the possibility of profitable future interactions may compel the informed agent not to expropriate the uninformed party and may therefore alleviate the problem of asymmetry.

## Information Production and Innovation

So far in this chapter, we have been mostly concerned with how information is interpreted in various economic settings. With few exceptions (Grossman & Stiglitz, 1980, being a notable one), the models discussed above say little about the actual process of acquiring new information. In most cases, the discussion of equilibrium starts by assuming the presence of private information. In this section, we intend to address the problem of information production.

The way information is produced and disseminated can have significant consequences for the economy. Sometimes changes in information production can trigger a total overhaul of incentive structures in a certain business field. The way credit rating agencies are compensated is one example: Until the 1970s, the agencies were paid by the users of financial information, while today most of them are being paid by the issuers of bonds these agencies rate. Many economists believe this creates conflicts of interest, but it may be interesting to note that the compensation structure was changed when photocopying technologies became widely available. In effect, disseminating ratings after obtaining them from the agencies became less costly. As a result, rating agencies found it difficult to privatize returns to their efforts. Other factors may have been responsible for the change, of course, but the availability of technology is likely to have played a role.

Since information is a public good, private production of information may sometimes be insufficient or even impossible. In cases when information is very valuable, the markets have found ways to cope with the problem by creating special structures and institutions that enable private information production. Financial intermediation discussed above is one example. Quite surprisingly, however, there are public institutions and mechanisms such as patent protection designed to safeguard private information.

Why is it necessary to create mechanisms that protect private information even as this may lead to information asymmetries and the associated problems? The answer is far more difficult than it may seem. We live in a world where scientific discovery and innovation drive technological progress. Many economists acknowledge that it is the rate of technological progress that stimulates economic growth more than any other factor. Since innovation is impossible without producing information, it is difficult to underestimate the importance of creating incentives to acquire information. The problem, however, is how much and what kind of information needs to be produced to ensure economic progress.

One may argue that due to the public-good nature of information, it is government-financed institutions that must perform most information production. Since benefits of information are often public, the public should be responsible for financing the production of this information. Even a cursory observation of economic realities, however, leads one to conclude that the share of private funding for research is far from trivial. The Organisation for Economic Co-operation and Development (2009) performs biannual surveys of science and technology indicators, which suggest that more than two thirds of research and development is performed by corporations. Apparently, publicly funded research may be associated with problems that go beyond the public-good nature of information.

One problem with public research is its remoteness from economic realities. Since the amount of information that can potentially be discovered is virtually unbounded, limited research funds and efforts must be allocated in a way that produces more or less tangible benefits. Aghion, Dewatripont, and Stein (2005) developed a theory to address the problem of research allocation between corporations and academic institutions. They assume that the primary difference between academic and private research is the degree of scientific freedom: In academia, a scientist may pursue whatever direction he or she finds interesting, while in the private sector, he or she will have to work on a specific project mandated by the management. Aghion et al. show that in this case, it is optimal to originate fundamental research in academia and delegate later stages to corporations (by later stages, they mean stages close to actual product implementation). They also show that there exists an optimal transfer point prior to which private sector research may inhibit innovation. Thus, the structure of incentives in social institutions may be suitable for some kinds of research and detrimental to others. Monopolistic information production within public entities is, therefore, suboptimal.

The debate about patent protection and whether it promotes or inhibits innovation is ongoing. This debate, however, is not the focus of our discussion. Our purpose is to show that it is impossible to create a single mechanism that would address all issues related to information production.

## Conclusion

Information economics is a growing field with high potential for future research. This chapter was intended to give an incomplete overview of the main themes in the literature and discuss their applicability to various economic problems. Our hope was to show that the structure of information environment has far-reaching consequences for economic behavior.

When information is asymmetric, a market breakdown can occur because of adverse selection and moral hazard: People will try to adjust their behavior to use the information available to them. Other agents anticipate this behavior and will either demand compensation or will refuse to participate in market transactions that will leave them at a disadvantage. Screening and signaling are two ways to deal with the problem of asymmetry. With signaling, the informed party behaves in a way that reveals private information. A credible signal can separate different types of informed agents: Some types of informed agents find signaling costly enough not to engage in it, while others find it profitable enough to send the signal. Screening is a device used by the uninformed party to extract information from informed agents. It translated into a menu of options, which are catered to different types of informed agents.

Screening, signaling, and information asymmetry have applications in various fields of economics. We discussed examples from the theory of credit rationing, corporate finance literature, insurance markets, literature on market efficiency, financial intermediation, and reputation. Each of these topics has important implications for policy debates.

Health care and medical insurance are one example. Some countries and states develop universal systems of health care, in effect requiring mandatory medical insurance from their residents. Sometimes such systems fail (as in some post-Soviet countries) while sometimes they seem to be very successful (as in France). The reasons why the results may be so different are not yet well understood, and the debate is likely to continue.

Private information and concerns about its use in financial markets led to continuous attempts on behalf of the governments to increase market transparency and design optimal systems of financial regulation. While progress has been made, economic crises of the past and the recent 2008 global financial meltdown are evidence that much still needs to be done. The structure of financial intermediation and the role of reputational capital are a major topic for academic research and policy debates. The development of e-commerce also stimulates reputation research since online identities are often easy to change, and reputation may be one of the ways to deal with unscrupulous deal making.

We also briefly discussed information production. This is a growing area of research, with information technology affecting every facet of our lives. We showed that it might be challenging to design optimal systems of information production. There are entire countries that place emphasis on public research, Russia being one example, while other governments favor a more laissez-faire approach to scientific discovery. The difference in these approaches may be in part responsible for why fundamental research succeeds in some places while applied research succeeds in others. Explaining and detailing such differences may be invaluable for future policy decisions.

There are many topics we had to leave untouched, such as the role of information in designing optimal compensation, dissemination of information through social networks, and many others. The student is referred to the list of suggested readings for further details on some of the topics.

## Note

1. In writing this section, I benefited from discussions and notes from Tom Chemmanur's corporate finance theory class at Boston College.

## References and Further Readings

Abreu, D., Pearce, D., & Stacchetti, E. (1990). Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica, 58,* 1041–1063.

Aghion, P., Dewatripont, M., & Stein, J. C. (2005). *Academic freedom, private-sector focus, and the process of innovation* (NBER Working Paper No. 11542). Cambridge, MA: National Bureau of Economic Research.

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

Akerlof, G. A., & Yellen, J. L. (1985). Can small deviations from rationality make significant differences to economic equilibria? *American Economic Review, 75,* 708–720.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy, 100,* 992.

Ellison, G., & Fudenberg, D. (1995). Word-of-mouth communication and social learning. *Quarterly Journal of Economics, 110,* 93–125.

Fama, E. F. (1965). The behavior of stock-market prices. *Journal of Business, 38,* 34.

Fama, E. F. (1970). Efficient capital markets. *Journal of Finance, 25,* 383–421.

Fama, E. F. (1991). Efficient capital markets: II. *Journal of Finance, 46,* 1575–1617.

Graham, J. R. (2000). How big are the tax benefits of debt? *Journal of Finance, 55,* 1901–1941.

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient prices. *American Economic Review, 70,* 393–408.

Hall, B. J., & Liebman, J. B. (1998). Are CEOs really paid like bureaucrats? *Quarterly Journal of Economics, 113,* 653–691.

Jensen, M. C., & Murphy, K. J. (1990). Performance pay and top-management incentives. *Journal of Political Economy, 98,* 225.

Leland, H. E., & Pyle, D. H. (1977). Informational asymmetries, financial structure, and financial intermediation. *Journal of Finance, 32,* 371–387.

Mikkelson, W. H., & Partch, M. M. (1986). Valuation effects of security offerings and the issuance process. *Journal of Financial Economics, 15*(1/2), 31–60.

Miller, M. H., & Rock, K. (1985). Dividend policy and asymmetric information. *Journal of Finance, 40,* 1031–1041.

Modigliani, F., & Miller, M. (1958). The cost of capital, corporation finance and the theory of investment. *American Economic Review, 53,* 433–443.

Myers, S., & Majluf, N. (1984). Corporate investment and financing decisions when firms have information that investors do not have. *Journal of Financial Economics, 13,* 187–221.

Organisation for Economic Co-operation and Development (OECD). (2009). *Main science and technology indicators* (Vol. 2008/2). Paris: Author.

Rajan, R. G., & Zingales, L. (1995). What do we know about capital structure? *Journal of Finance, 50,* 1421–1460.

Rothschild, M., & Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics, 90,* 629–649.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics, 87,* 355–374.

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy, 69,* 213.

Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *American Economic Review, 71,* 393–410.

Titman, S., & Wessels, R. (1988). The determinants of capital structure choice. *Journal of Finance, 43,* 1–19.

# 72

# Forensic Economics

Lawrence M. Spizman
*State University of New York at Oswego*

The National Association of Forensic Economics (NAFE) defines forensic economics as "the scientific discipline that applies economic theories and methods to the issue of pecuniary damages as specified by case law and legislative codes" (National Association of Forensic Economics, n.d.-a). During the litigation process, economists determine the value of economic damages, testify, and critique the opposing experts' economic analysis.

The purpose of this chapter is to provide a general overview of the economic issues in a typical personal injury and wrongful death litigation tort, which is a private or civil wrong. The chapter also discusses the ethical issues involved when using different methodologies to estimate damages.

Economic damages typically are presented to the trier of fact (which can be a jury or judge) as the last phase of a trial. The importance of this should not be underestimated. Before there can be any award for economic damages, the plaintiff (the one suing for damages) must show that the defendant (the one being sued) is liable. If an attorney offers a meticulous case for liability but the economist presents unrealistic damage estimates, then the initial favorable impression of the trier of fact may be reversed if the opinions of the last expert to testify are perceived as bogus.

Before a trial, the plaintiff and defendant participate in settlement negotiations. During these negotiations, the plaintiff will present the demands for economic damages while the defendant might have a counteroffer. The plaintiff will usually retain an economist to estimate the economic damages. The plaintiff's economist will be identified and will have to present a report showing how the results were derived. In federal cases as well as in some state jurisdiction cases, the economists may be deposed and asked a series of questions to find out how damages were estimated. This information will later be used at a trial. The defense may or may not list an economist as an expert to counter the claims of the plaintiff. However, the defense often will retain a consulting expert to help prepare and critique the plaintiff's economic analysis. The defense does not disclose who the consulting expert is, nor does the consultant testify. The plaintiff's economic expert will not know if there is a defense expert critiquing and checking the results of his or her analysis. If the case settles before trial, the expert's work is complete. Even though most cases settle before trial, it is prudent for the testifying expert to estimate damages assuming that a trial will occur. A testifying expert is sworn to tell the truth. Therefore, the economic analysis must be based on accurate information even if there is only a remote chance that a trial will occur. Spizman (1995) discussed the negotiating strategy process between the plaintiff and defendant given the small probability of a trial occurring.

Forensic economics as a formal academic discipline began in 1986 with the formation of NAFE. NAFE started publishing the first journal devoted exclusively to forensic economics, the *Journal of Forensic Economics,* in 1988. NAFE also published the journal *Litigation Economic Digest* for several years before it ceased publication. In addition, NAFE sponsors sessions devoted to forensic economics at the major economic conferences. The American Academy of Economics and Financial Experts started publishing the *Journal of Legal Economics* in 1991. The American Rehabilitation Economics Association began publishing *The Earnings Analyst* in 1998.

Arguably the most important book in forensic economics is by Martin (2009). First published in 1988, it had 21 annual supplements and includes special sections written by more than 40 leading forensic economists.

Kaufman, Rodgers, and Martin (2005) published a compilation of major articles dealing with personal injury or wrongful death. Ireland et al. (2007) and Brookshire, Slesnick, and Ward (2007) discussed the major issues of forensic economics.

## Ethics and Assumptions of Damage Models

The discipline of forensic economics is unique in economics because most academic forensic economists are also consultants to the legal community. Since forensic damage models are based on assumptions that may favor one side in litigation, practitioners are confronted with ethical issues dealing with the impact of their models' assumption on litigants. Consequently, forensic economists must go beyond the simplifying assumptions made in introductory economics classes and understand the consequences of the models' assumptions. Because experts are not advocates for either the plaintiff or defendant (attorneys are), it is crucial that their assumptions be consistent and not change depending on which side retains them. Neutrality can be difficult to maintain when the marketplace rewards those providing opinions beneficial to the retaining side. The ethical consistency dilemmas are real and not abstract. Different sections in this chapter will address these ethical consistency issues. However, it is important to realize that each case is unique and that research and new data may warrant changing methodology on specific issues. Changes must be defended if they differ from past practices.

## Law

Each state, as well as the federal government, has different laws pertaining to estimating economic damages. Nevertheless, the methodology of estimating damages within the legal parameters is remarkably consistent from one jurisdiction to another. The purpose of estimating damages is to restore the plaintiff's economic condition to what it was prior to the tort, or to make the plaintiff whole.

## Life, Work Life, and Healthy Life Expectancy

Each component of damages depends on how long the loss lasts. Lost earnings depend on work life expectancy, which is the number of years the plaintiff would have worked if not injured or deceased. Long-term health care resulting from an injury and pension losses depend on the plaintiff's life expectancy. Other losses such as household services last as long as the plaintiff is healthy enough to provide those services. Skoog and Ciecka (2003) provide work life expectancy tables, while Arias (2005) generates life expectancy tables. *Healthy Life Expectancy* (Expectancy Data, 2010) provides tables for healthy life expectancy. All these tables are broken down by various demographic characteristics such as age, gender, race, and educational levels.

Because all expectancy tables are based on the age of the plaintiff, an important issue in determining expectancies is whether to use the plaintiff's age on the date of the incident or on the date of the trial. Using the age on the date of trial extends an individual's work life and life expectancy beyond what it would be if using the date of the incident. Thus, the number of years the loss continues is increased. Damages should commence based on the plaintiff's age on the date of the incident rather than on the date of trial, unless the laws of a specific jurisdiction require otherwise. Ethical consistency requires that the forensic economist not choose one starting date for the plaintiff in order to have a longer life or work life expectancy and another starting date for the defense to get lower expectancies.

Although the trial date is generally not used to determine expectancies, it is used to delineate past losses from future losses. This is important because future losses, not past losses, are discounted to present value.

### Life Expectancy

Life expectancy is the number of years an average person would have lived but for the tort. Life expectancy is relevant if the plaintiff requires lifetime medical care and has a lifetime-defined benefit pension plan. Pain and suffering, which is awarded by the trier of fact and not calculated by an economist, also can be awarded for life. Arias (2005) presents life expectancy by age, gender, and race. The use of race-neutral tables to determine life expectancy is often required by case law. If the law does not specify race-neutral life expectancy tables, then the forensic economist should not choose to use race tables when those tables favor one side and use race-neutral tables when it favors another side. For example, suppose black males have lower life expectancies than white males and the all-male life expectancy category is what is normally used. The black male category should not be used when the plaintiff is black in order to get a lower life expectancy if the defendant is estimating damages. If the black male was a physician, would his life expectancy be any different than a white male physician? Switching life tables to benefit one side over the other raises the ethical consistency issue.

The plaintiff's age on the date of the incident should be used to determine life expectancy. Suppose a female was 35 at the time she was involved in an accident and the trial occurred 3 years after the accident. The correct work life should be for a 35-year-old female. Using the work life for a 38-year-old female would add more years to the work life. The probability of the plaintiff living from 35 to 38 is

100% since she is already 38 years old. If losses continue for 30 more years, adding one additional year to work life can significantly increase losses when the effects of 30 years of compounding damages are considered. The last extra year will be the largest yearly loss. Ethical consistency requires using the same methodology for the plaintiff and defendant.

Another potential problem in estimating damages occurs with partial years of loss. The first and last year loss is usually a partial year unless the tort commenced on January 1 or ended on December 31. For example, if an injury occurred May 4, 2009 (2009.34), the first year's loss is only 66% of the year since 34% has already occurred. If the loss continues until the plaintiff's work life expectancy year 2045.3, then the loss is only for 30% of the year 2045. Sometimes, the first year and last year are rounded to complete years with the false claim that they offset each other. The final year's loss is the highest because of growth and compounding, and the first year's loss is the lowest, and thus they cannot offset each other. Rounding to full years in one case and using partial years in another to get a loss favorable to either the defendant or plaintiff would be ethically inconsistent.

### Work Life Expectancy

Smith (1982) used the increment-decrement Markov model, which considers the probability of an individual's movement from being active to inactive in the labor force. Smith (1986) updated these tables using 1979 data. The Department of Labor stopped publishing work life tables but continues providing data through the Current Population Survey, allowing economists to keep work life tables up-to-date. One of the most recent tables was published by Skoog and Ciecka (2003).

Work life tables show the number of years, on average, a person will be working or actively looking for work throughout his or her life. The tables do not tell us when an individual retires from the labor force. For example, a 42-year-old female with a bachelor's degree who is currently employed has, on average, 19.03 years of working or actively looking for work for the remainder of her life. Her work life is to age 61.03. This does not mean she will retire at age 61.03; rather, it tells us the number of continuous years she can be expected to either work or look for work. The tables take into account a worker's being out of the labor force for various reasons. In essence, work life frontloads the person's remaining years in the labor force because she may still be working past her work life. Thus, if a female is out of the labor force for childrearing purposes, that is factored into her work life expectancy. For example, if a worker leaves the labor force to get a degree in business administration or is injured temporarily and later returns to work, then that is factored into the work life tables.

To properly use work life tables, you need to know whether the plaintiff was active or inactive at the time of the injury. Age, gender, and levels of education also determine work life. Because work life tables frontload the loss, some forensic economists use the life, participation, and employment (LPE) method. This method takes the probabilities of participation in the labor force, survival, and unemployment to determine the expected value of future earnings.

### Healthy Life Expectancy

Some damages such as household services may not continue for life because as a person's health normally deteriorates with age, the amount of household services he or she is able to perform diminishes. To account for this deterioration, *Healthy Life Expectancy* (Expectancy Data, 2010) publishes tables showing the number of years a person considers his or her health to be excellent without any limitations to activities. Even though a person's health may decline, he or she is still capable of performing household services. *Healthy Life Expectancy* also has tables of full-function life expectancy (FFLE), which has fewer years than life expectancy but more than healthy life expectancy (HLE). The ethical consistency issue requires the use of either HLE or FFLE to determine the number of years household services would have continued but for the tort. Using one table for the plaintiff and another for the defendant would be inconsistent and therefore unethical.

## Wage and Salary Loss

Past and future earning losses resulting from an injury or death are recoverable. Legal parameters determine whether expected earnings (earnings the plaintiff expected to earn prior to the tort) or earning capacity (earnings the plaintiff had the ability to earn prior to the tort) are to be used. Upon establishing preinjury earnings, postinjury earnings (residual earnings) have to be determined and deducted from lost earnings. In wrongful deaths, there are no residual earnings. Vocational rehabilitation experts usually determine the future earnings potential given the impaired condition of the plaintiff.

Since 1040 tax forms can show additional income (such as spouse's income, business income, and royalties), the best sources for preinjury earnings are W-2 tax forms. The Social Security Administration also provides yearly earnings that can easily be obtained if the plaintiff's W-2 forms are unavailable.

When the plaintiff's work history is well documented, it is easier to establish the base earnings necessary to estimate future earning losses. If earnings vary from year to year, average earnings can be used to determine the base earnings. It is debatable how many prior years to use in establishing the average, but 3 to 5 years should be appropriate. Past earnings should be in current dollars (constant dollar equivalents) before taking an average. For example,

if using average earnings from 2003 through 2008, then 2003 earnings should grow by the rate for 5 years, and 2004 earnings grow by 4 years, 2005 by 3 years, and so forth. The average is based on the current dollar earnings.

Earning capacity can be used in the absence of earning records. Earning capacity considers those occupations that an individual is capable of entering. Earnings for broad occupational categories are found in National Occupation and Wage Estimates (U.S. Department of Labor, 2007b) and State Occupational Employment and Wage Data (U.S. Department of Labor, 2007c).

If the plaintiff is an injured child or a recent graduate, then a broader category of average earnings is required. One such category is educational attainment. Wages for different age, gender, and race cohorts can be found in the U.S. Census Bureau (2007) and Expectancy Data (2008, 2009, 2010).

When broad statistical averages are required, then median (not mean) earnings should be used. For example, if 9 of 10 workers in the sample earn $50,000 and one earns $250,000, mean earnings are $70,000. One high-salaried worker skews the average upward. If the sample size is small, this becomes all the more important. Median earnings are $50,000 because half the workers earn more and half earn less. Since median earnings are less than mean earnings, the ethical consistency issue requires that the mean not be used if retained for the plaintiff and the median if retained for the defense.

Union or professional dues should be deducted from lost earnings because they are a cost to the plaintiff's job maintenance. Even though union dues are small, ethical consistency issues require that you be consistent when choosing to deduct union dues.

## Growth of Earnings

Once the plaintiff's current wage or salary is established, the future growth rates of wages have to be determined. If past employment records are available, then average past growth rates can be used to project future increases. When employment records are not available or suitable, then the use of general wage or price indexes such as the Consumer Price Index (CPI) or Employment Cost Index (ECI) can be used for estimating wage growth rates. The implied assumption when using the CPI to determine wage growth rates is that the plaintiff's wages will only increase by the level of the CPI, which shows the increase in the price level for goods and services. An index that gives a broader measure of wage and salary increases is the ECI. The ECI not only shows the historical growth rate for all workers but also allows tailoring the growth rates to a specific industry group. In addition to showing straight time salary and wage rates, the ECI includes earning incentives, cost of living adjustments, and production and earning bonuses. In addition, the ECI has an index for employee-provided benefits.

The number of past years required to establish the average growth rate is not universally agreed upon and should be justified. However, to avoid the ethical consistency issue, the same number of years to estimate the average should be used for both plaintiffs and defendants rather than using one number that provides a higher growth rate for the plaintiff and then using a different number that provides a lower growth rate for the defense.

The compounding of growth rates can magnify the total value of losses when there is what appears to be a small percentage difference. The longer the timeframe growth is, the greater the loss.

## Fringe Benefit Losses

Employer-provided fringe benefits that are lost due to the injury have value that the plaintiff or the plaintiff's family can recover. It is important not to double count fringe benefit losses. For example, if the plaintiff received 3 weeks of paid time off annually and the plaintiff is already being compensated for 52 weeks of lost earnings, then including another 3 weeks of earnings for time off would be double counting the loss.

The two largest components of fringe benefit losses are often health insurance and retirement benefits. When employee fringe benefits are well established, then those benefit amounts should be used as a basis for lost fringe benefits. When it is unclear what the benefits are or if fringe benefits have not been established, then statistical averages of fringe benefits as a percentage of total earnings can be used.

One data source that provides fringe benefits as a percentage of earnings is *Employer Costs for Employee Compensation* (U.S. Department of Labor, 2008). Growth rates for fringe benefits can be determined from the *Employment Cost Index* (U.S. Department of Labor, 2009a).

*Employee Benefits in Private Industry* (U.S. Department of Labor, 2007a) provides information on the participation of workers receiving different types of fringe benefits and the frequency of benefit use. The *Employer Health Benefits Annual Survey* (Kaiser Family Foundation, 2009) provides information about the cost of employer-provided health insurance. Some state health insurance departments provide health cost by local jurisdictions.

If general statistical data are used, it is important to be familiar with both the data source and the actual data. For example, the *Employer Costs for Employee Compensation* (U.S. Department of Labor, 2008) tables data show the percentage of fringe benefits to the total compensation package. However, the total compensation package already includes fringe benefits so the percentage provided in the data should not be used. Instead, the economist should calculate fringe benefits as a percentage of wages and salaries and use that amount in estimating damages. The tables provide different categories of fringe

benefits so for each case the appropriate benefit can be used. Another source of general data of fringe benefits is the U.S. Chamber of Commerce's annual *Employee Benefits Study*. However, there are many statistical issues about the Chamber of Commerce study that should raise red flags about its use. Spizman (2008) suggests not using this study because of its bias and extremely small self-selected sample size.

## Health Insurances

There are several ways to calculate the loss to the plaintiff or the plaintiff's family for medical insurance. No one method is absolutely correct, so the best method depends on the facts of each case. The first method is to award the plaintiff the cost of the employer's medical insurance premium. However, this may not allow the plaintiff to purchase comparable insurance because an employer's group rates are often lower than individual rates. A second approach is to get price quotes for health insurance to replace the coverage for comparable medical insurance. Some states' insurance departments provide comparable rates by state regions. Since employer-provided insurance benefits are well defined, it is critical to find the replacement value for similar coverage. Online price quotes simplify the process of getting costs. However, often online policies provide minimum coverage that would not make the plaintiff whole.

A third method of estimating insurance costs is used when it is not clear whether the plaintiff received or would have received medical insurance in the future. For example, when a minor child or a recent graduate is injured before entering the labor market, what value should be placed on lost health insurance? If a worker has an entry-level job without medical insurance but future employment opportunities would provide health insurance, how is that future insurance valued? Given these circumstances, general statistical data that take the average percentage cost of health insurance may be appropriate to use in estimating lost health insurance. Several issues should be considered when using statistical averages. Suppose that, on average, health insurance is 12% of wages and salaries. Consequently, a worker earning $30,000 a year with family coverage will be allocated $3,600 a year ($300 per month) for health insurance losses. If the replacement cost is $12,000 a year, then the plaintiff is undercompensated by $8,400. If the plaintiff is an executive receiving the same coverage for the same price but earning $200,000 a year, then losses would include $24,000 a year for health insurance, thus overcompensating the plaintiff by $12,000. Health insurance is a quasi-fixed labor cost and should be valued the same for all employees; that is, it is fixed per worker no matter how much workers earn or how many hours a week they work.

Regardless of how health insurance losses are determined, the employee's preinjury contribution to health insurance should be deducted from any loss because that is an expense before the plaintiff incurred the loss.

## Pension Loss

There are two types of pensions that workers usually participate in: a defined contribution and a defined benefit. A defined contribution is a percentage amount that an employer contributes to the employee's 401k or similar plan. The employer's contribution is the loss to the plaintiff. If an employee contributes to his or her own retirement plan, then that contribution should not be counted as a loss because the percent contribution is already being replaced from earnings; to replace it again would double count the loss. Since many employee pension contributions are tax deferrals, it is important to use Medicare earnings rather than Social Security reported earnings. Medicare earnings are higher because Social Security earnings are reduced by pension deferrals.

A defined benefit is often based on a formula that takes the number of years of employment multiplied by a final average salary multiplied by some percentage amount. Any pensions that accumulated before the injury should be deducted from the pension the plaintiff would have accumulated if not for the injury. For example, if a plaintiff had not been injured and would have received a monthly $1,500 pension but instead, as a result of the injury, only received a $400 pension monthly, the loss is $1,100 a month. If a plaintiff starts receiving the $400 a month pension before his or her normal retirement age, that amount is part of the offset against the full pension he or she would have received if not injured.

Other types of fringe benefits losses that may be considered are premiums paid by the employer on a life insurance policy and the use of a company car or cell phone for personal use.

## Social Security and Fringe Benefits

Federal Insurance Contributions Act (FICA) taxes are divided between employees and employers, with each paying 7.65% of the employees' salaries. The Social Security benefits portion of FICA is 6.2% on the first $106,800 (2009 rates) of an employee's income. This includes 5.3% for Old Age Survivors and .9% for disability insurance. The Medicare tax portion of FICA is 1.45% of every dollar of earnings, with no limit on earnings. Whether or not Social Security benefits should be included as part of fringe benefits losses is a controversial issue. If a claim is being made for lost Social Security, then the 5.3% for Old Age Survivors should be used, not the full 7.65% employer contribution. Including Medicare and disability benefits as a loss would be double counting since they would be forthcoming to the injured plaintiff. One reason many economists do not include FICA taxes as part of lost fringe benefits is because taxes are ignored in most jurisdictions.

A more compelling reason not to include the plaintiff's portion of FICA as a loss is that the plaintiff no longer has to make a matching contribution to FICA, which offsets the employer's contribution.

Rodgers (2000) shows that using a percent loss for Social Security is a poor estimate of lost Social Security benefits. If there are any losses of Social Security, in most circumstances they are small. However, it is important to recognize that the circumstances of each case (e.g., an injured young child) can alter the approach used in determining whether FICA should be included as a part of fringe benefits losses.

## Household Service Losses

Household services are those activities performed outside the paid marketplace that have pecuniary value that can be quantified by an economist. Household services provided by the plaintiff that he or she is no longer capable of doing because of the accident are economic damages. Household services may include cleaning, cooking and cleaning up, doing laundry, shopping, maintaining the home and vehicles, managing the household, providing transportation for the household, caring for children, and other types of services unique to the plaintiff.

Household services are not a loss if the plaintiff never performed them in the past and if there is no evidence to support that he or she would have performed them in the future. A method to determine the pre- and postinjury hours of household services is to have the plaintiff fill out a questionnaire asking how many hours were spent doing specific household work before and since the accident. The difference between the two is the reduction of household services, which would then be valued. The potential for self-reporting bias has to be recognized. However, the purpose of the survey is to show that a foundation for losses does exist. Without a foundation, it may be difficult to claim lost household services.

The *Dollar Value of a Day* (*DVD;* Expectancy Data, 2009) provides general statistical averages that estimate the hours of household services performed categorized by age, number of children, marital status, and gender. The *DVD* relies on the latest government data and is widely used. The DVD provides national average hourly wages for household services with adjustments for different geographic areas within each state.

It is important to remember that the role of the economist is to provide guidance to the trier of fact with respect to valuing losses. The trier of fact can increase or decrease losses based on the testimony of the plaintiff or the plaintiff's survivors. The *DVD* provides an excellent starting point for determining the number of hours of household services. While there are other sources and methods for computing household services, ethical consistency requires that the economist not choose one source for the plaintiff and another for the defense. Consistency and neutrality are important.

### Healthy Life

The aging process limits a person's ability to perform the same level of household services when older compared to when he or she was younger and healthier. Consequently, it may not be appropriate to use a person's life expectancy to project future losses of household services. The publication *Healthy Life Expectancy* (Expectancy Data, 2010) considers the diminution of household services because of aging by providing tables showing how many remaining years of healthy life expectancy an individual has based on age, race, and gender. The publication also provides full-function life expectancy tables, anticipating that household activities will be reduced rather than eliminated when an individual is sick.

## Personal Consumption Deduction

Wrongful death requires that a deduction be made for the income the decedent would have used for his or her personal consumption of items such as food, clothing, and personal care. Some states allow for personal maintenance rather than personal consumption. Personal maintenance is the amount that would have been spent by the decedent to maintain himself or herself to attain his or her earning capacity. The percent deduction for the decedent's personal consumption can be found in Ruble, Patton, and Nelson (2007). The percentage deduction changes as family size and income change.

There is some difference of opinion as to whether fringe benefits should be reduced for personal consumption since the survivors would not have received the dollars the deceased would have spent on maintenance had he or she lived. Consequently, as a result of the plaintiff's death, the survivors did not lose these dollars. A personal consumption deduction may or may not be appropriate when dealing with household services. If the trier of fact determines that the decedent's household services cannot be split between members of the household, then it may not be appropriate to deduct personal consumption from household services. One solution is to estimate lost household services in death cases both with and without the personal consumption deduction.

There are situations when personal consumption deductions are made for personal injury. For example, a severely injured person may have to spend the rest of his or her life in an extended-care facility that provides all services that the plaintiff formerly performed for himself or herself. If there is a claim for damages for the cost of the extended-care facility, then to count household services losses (which are being provided by the facility) would be double counting that loss.

## Present Value

Since future dollars are worth less today, future losses must be discounted to present value. Each jurisdiction has different rules about discounting. Some states do not discount, while others specify what the discount rate should be. Because of the inverse relationship between discount rates and present values, choosing the appropriate discount rate can be a very contentious issue. That being said, what is widely accepted is that low risk and safety should be the guiding principle in choosing discount rates. This usually means U.S. government bonds. Kaufman et al. (2005) present the different approaches used to determine discount rates. Historical averages of some instruments, spot rates at the time of the analysis, short-term versus long-term rates, and net discount rates (the difference between the growth rate and the discount rate) can also be used. Because small differences in the discount rate (as well as growth rates) can affect total damages, justifying the rate being used for discounting is important. Bloomberg.com (n.d.) and *Federal Reserve Statistical Release* (n.d.) provide current rates.

Ethical consistency issues become most noticeable when determining the discount rate because lower discount rates favor the plaintiff while higher rates favor defendants. Whether retained by the plaintiff or defendant, the same discount rate should be used. It is considered unethical to choose a low rate for the plaintiff and a high rate for the defendant. Since court testimonies are a matter of record and prior economic reports are often seen and saved by opposing experts, it is not difficult to find inconsistencies in methodological approaches between plaintiff and defense cases that are not looked upon favorably by the courts.

## Life Care Plan

Catastrophic injuries and disabilities to children or adults often require lifetime health and personal care. Life care planners provide plans that show the different components of care that the plaintiff will require in the future. Examples are prescription and nonprescription drugs, future medical care, and future attendant care. The life care plan presents the required care, the length of care, the frequency of care, and the current cost of the care. Once this information is provided, the economist then estimates the future cost of the life care plan. Each component of the life care plan grows by the appropriate inflation rate associated with the historical average from the Medical Care Price Index subsection of the CPI. Communications between the life care planner and economist to match the life care component to the medical care index component are useful. Items from the life care plan such as transportation are matched to the appropriate transportation index of the CPI. Double counting between the life care plan and other elements of damages must be avoided. For example, if the life care plan provides for all future medical care, then lost medical insurance should not be an element of damages. If a special van is required for transportation, then only the additional cost of modifying the van is a loss and not the full cost of the van, since the plaintiff would have bought transportation regardless of the accident. If the life care plan includes funding for certain household services, care must be taken not to double count lost household services. If the jurisdiction where the trial is occurring requires discounting, the present value of the life care plan should be made.

## Collateral Source Payments

Collateral sources are payments the plaintiff receives due to the injury from a third party from insurance or the government. Collateral sources are ignored and not deducted from any loss if the payment is made by an entity that is not the defendant. The reason is that the third party who is paying the plaintiff can have a lien on any recovery from the lawsuit. Thus, if that amount is deducted from the loss, the remaining award will not be large enough to cover damages, and the defendant will benefit because he or she has to pay less in damages. However, if the defendant paid for insurance, then he or she is entitled to deduct that amount from the award. Collateral sources are often a confusing aspect of damages, and legal guidance may be required.

## Taxes

Deducting federal and state income taxes from any estimated loss depends on the jurisdiction in which the case is being tried. Taxes are generally ignored in most states and are deducted in federal jurisdictions. Nevertheless, it is important to find out how the court's jurisdiction treats taxes.

## Conclusion

This chapter addressed the key issues of forensic economics while stressing the interrelationships between law and economics when dealing with wrongful death and personal injury litigation. Because forensic economics is intricately tied to changes in both statutes and common law, there is tremendous potential for future research and growth in the area. Many ethical dilemmas confronting a practitioner were highlighted. Most of the topics discussed in this chapter are developed more fully in Martin (2009) and the book of readings by Kaufman et al. (2005).

## References and Further Readings

American Academy of Economic and Financial Experts. (n.d.). *Journal of Legal Economics*. Index. Available at http://www.econ.unt.edu/jle

American Rehabilitation Economics Association. (n.d.). *The Earnings Analyst.* Available at http://www.a-r-e-a.org/journal.shtml

Arias, E. (2005). United States life tables. *National Vital Statistics Reports, 58*(10). Available at http://www.cdc.gov/nchs/data/nvsr/nvsr56/nvsr58_10.pdf

Bloomberg.com. (n.d.). *Government bonds.* Available at http://www.bloomberg.com/markets/rates/index.html

Brookshire, M., Slesnick, F., & Ward, J. (2007). *The plaintiff and defense attorney's guide to understanding economic damages.* Tucson, AZ: Lawyers & Judges Publishing.

Expectancy Data. (2008). *Full-time earnings in the United States: 2007 edition.* Shawnee Mission, KS.

Expectancy Data Inc. (2009). *The dollar value of a day: 2008 valuations.* Shawnee Mission, KS: Author.

Expectancy Data Inc. (2010). *Healthy life expectancy: 2004 tables.* Shawnee Mission, KS: Author.

*Federal Reserve Statistical Release: Selected Interest Rates (Daily).* (n.d.). Available at http://www.federalreserve.gov/Releases/H15/update

Ireland, T., Horner, S., Rodgers, J., Gaughan, P., Trout, R., & Piette, M. (2007). *Expert economic testimony: Reference guides for judges and attorneys.* Tucson, AZ: Lawyers & Judges Publishing.

Ireland, T., & Ward, J. (2002). *Assessing damages in injuries and deaths of minor children.* Tucson, AZ: Lawyers & Judges Publishing.

Ireland, T. R., & Depperschmidt, T. O. (1999). *Assessing family loss in wrongful death litigation: The special roles of lost services and personal consumption.* Tucson, AZ: Lawyers & Judges Publishing.

Kaiser Family Foundation and Health Research Family Trust. (2009). *Employer Health Benefits Annual Survey.* Available at http://ehbs.kff.org/pdf/2009/7936.pdf

Kaufman, R., Rodgers, J., & Martin, G. (Eds.). (2005). *Economic foundations of injury and death damages.* Cheltenham, UK, Northampton, MA: An Elgar Reference Collection.

Martin, G. (2009). *Determining economic damages.* Costa Mesa, CA: James Publishing.

National Association of Forensic Economics. (n.d.-a). Home page. Available at http://nafe.net/default.aspx

National Association of Forensic Economics. (n.d.-b). *Journal of Forensic Economics.* Back issues. Available at http://nafe.net/JFE/Forms/AllItems.aspx

National Association of Forensic Economics. (n.d.-c). *Journal of Forensic Economics.* Index of articles. Available at http://www.econ.unt.edu/jle/JLEBackissues19912008.pdf

National Association of Forensic Economics. (n.d.-d). *Litigation Economic Digest.* Back issues. Available at http://nafe.net/LER/Forms/AllItems.aspx

Rodgers, J. (2000). Estimating loss of Social Security benefits. *The Earnings Analyst, 3,* 1–28.

Ruble, M., Patton, R., & Nelson, D. (2007). Patton-Nelson personal consumption tables 2005–2006. *Journal of Forensic Economics, 20*(3), 217–225.

Skoog, G. R., & Ciecka, J. E. (2003). The Markov (increment-decrement) model of labor force activity: Extended tables of central tendency, variation, and probability intervals. *Journal of Legal Economics, 11,* 23–87.

Smith, S. (1982). *Tables of working life: The increment-decrement model* (Bulletin 2135). Washington, DC: Bureau of Labor Statistics.

Smith, S. (1986). *Work life estimates: Effects of race and education* (Bulletin 2254). Washington, DC: Bureau of Labor Statistics.

Spizman, L. (1995). The defense economist's role in litigation settlement negotiations. *Journal of Legal Economics, 5*(2), 57–65.

Spizman, L. (2008). Sample selectivity bias of the U.S. Chamber of Commerce employee benefits study. *The Earnings Analyst, 10,* 49–63.

U.S. Census Bureau. (2007). *Educational attainment—people 18 years old and over, by total money earnings in 2007: Work experience in 2007, age, race, Hispanic origin, and sex* (PINC-04). Available at http://pubdb3.census.gov/macro/032008/perinc/toc.htm

U.S. Department of Labor, Bureau of Labor Statistics. (2007a). *Employee benefits in private industry.* Available at http://www.bls.gov/ncs/ebs/sp/ebsm0006.pdf

U.S. Department of Labor, Bureau of Labor Statistics. (2007b). *National occupation and wage estimates.* Available at http://www.bls.gov/oes/current/oes_nat.htm

U.S. Department of Labor, Bureau of Labor Statistics. (2007c). *State occupational employment and wage data.* Available at http://www.bls.gov/oes/current/oessrcst.htm

U.S. Department of Labor, Bureau of Labor Statistics. (2008). *Employer costs for employee compensation.* Available at http://stats.bls.gov/schedule/archives/ecec_nr.htm and http://www.bls.gov/news.release/pdf/ecec.pdf

U.S. Department of Labor, Bureau of Labor Statistics. (2009a). *Employment cost index.* Available at http://data.bls.gov/PDQ/outside.jsp?survey=ci

U.S. Department of Labor, Bureau of Labor Statistics. (2009b). *Employment cost index compensation component.* Available at http://data.bls.gov/PDQ/outside.jsp?survey=ec

U.S. Department of Labor, Bureau of Labor Statistics. (n.d.). *Consumer Price Index: All urban consumers* (Current Series). Available at http://data.bls.gov/PDQ/outside.jsp?survey=cu

# 73

# Economics of Crime

Isaac Ehrlich

*State University of New York at Buffalo*

The persistence of "crime" in all human societies and the challenges it imposes for determining how to enforce laws enjoining it have attracted the attention of scholars throughout human history, including, in particular, utilitarian philosophers and early economists such as Beccaria, Paley, Smith, and Bentham. Indeed, in view of its empirical regularity, sociologists such as Durkheim adopted the view that "crime, in itself, is a normal social phenomenon," rather than an aberration of human nature. It was not until the late 1960s, especially following the seminal work by Becker (1968), however, that economists reconnected with the subject in a systematic fashion, using the modern tools of economic theory and applied econometrics.

The essence of the economic approach, as restated by Becker (1968), lies in the assumption that potential offenders respond to incentives and that the volume of offenses in the population can therefore be deterred or prevented through an optimal allocation of resources to control crime. The objective of social policy is specified as minimization of the aggregate social loss from crime and law enforcement. Based on this criterion, Becker derived a comprehensive set of behavioral propositions and optimal enforcement strategies involving the certainty of apprehending and convicting offenders, the severity of punishment to be imposed on those convicted, and the selection of optimal instruments of punishment. The "deterrence hypothesis," as stated by Ehrlich (1973, 1975b, 1982, 1996), expands the scope of the relevant incentives by which offenders can be motivated or controlled. It highlights the relative roles and limitations of "negative" incentives such as the prospect of apprehension relative to conviction and punishment, whether by public law enforcement or private self-protection efforts, as well as "positive" incentives such as opportunities for gainful employment for workers at the lower end of the earnings distribution or education and rehabilitation efforts as deterrents to criminal activity. In this approach, the analysis of crime shares some formal similarities with that of occupational choice in labor-theoretic settings.

For this approach to provide a useful approximation to the complicated reality of crime, it is not necessary that all offenders respond to incentives, nor is the degree of individual responsiveness prejudged; it is sufficient that a significant number of potential offenders so behave on the margin. By the same token, the theory does not preclude a priori any category of crime, as offensive or heinous as it may be, or any class of incentives. Indeed, economists have applied the deterrence hypothesis to a myriad of illegal activities, from tax evasion, corruption, and fraud to robbery, murder, and terrorism.

## Theory

The economic approach to criminal behavior can be summarized by the following syllogism: "People respond to incentives. Offenders are people too. Therefore, offenders respond to incentives." Crime, in turn, inflicts material and emotional harm on both individual victims and on society as a whole and disrupts the foundations of civil society and efficient resource allocation. This is true even in the case of petty theft, which entails just a small redistribution of wealth from victim to offender. Theft involves a net social loss by virtue of the fact that thieves spend their time and energy on effecting a redistribution of wealth instead of creating new wealth. In addition, individuals and society spend resources on protection of property and avoidance of

emotional loss, as well as the potential loss of life and limb from being victims of serious crime, which is another significant drag on the economy and the pursuit of happiness. Therefore, theft, let alone more serious crime, entails a significant social cost, and society has a strong incentive to resist it in various forms.

In Becker's (1968) analysis, equilibrium volume of crime reflects the interaction between offenders and the law enforcement authority, and the focus is on optimal probability, severity, and type of criminal sanction—the implicit "prices" society imposes on criminal behavior to minimize the aggregate income loss from crime. This thesis has powerful implications concerning the choice of an optimal level of resources to be devoted by society to combat crime, as well as the optimal combination of law enforcement instruments to be imposed—the probability of apprehending and convicting offenders, the magnitude of the punishment to be meted out for crimes of different severity, and the form of the sanction to be imposed: imprisonment or monetary compensation (see Becker, 1968; Stigler, 1970).

Subsequent work has focused on more complete formulations of components of the criminal justice system, especially the supply of offenses, the production of specific law enforcement activities, and alternative social welfare criteria for producing optimal law enforcement strategies. Another methodological evolution has expanded the basic analytical setting of the Becker (1968) model by addressing the interaction between potential offenders (supply), consumers and potential victims (private "demand" for illegal goods or "derived demand" for protection), and deterrence and prevention (government intervention). This "market model" applies not just in the case of transactions involving direct demand for illegal goods and services, such as illicit drugs and prostitution, but also theft, robbery, and murder, for which the "demand" side derives from the private demand for individual safety. In this setting, government intervention works as a form of both demand and supply management, which can in principle combine elements of pure deterrence, such as monetary fines, with methods of individual control, such as incapacitation (retention, imprisonment, or confiscation of illegal goods) and rehabilitation of convicted offenders. This virtual market for offenses (Ehrlich, 1981, 1996) has later been extended to include interactions of crime with the general economy as well, including the prospect that crime could harm creative economic activity and thus economic growth and development. These extensions are discussed in greater detail in the following sections. For specific articles on which the following discussion is based, see Ehrlich and Liu (2006).

## Supply

The extent of participation in crime is generally modeled as an outcome of the allocation of time among competing legitimate and illegitimate activities by potential offenders acting as expected-utility maximizers. While the mix of pecuniary and nonpecuniary benefits varies across different crime categories, which attract offenders of different attitudes toward risk and proclivities ("preferences") for crime, the basic opportunities affecting choice are identified in all cases as the perceived probabilities of apprehension, conviction, and punishment, the marginal penalties imposed, and the differential expected returns on competing legal and illegal activities. Entry into criminal activity and the extent of involvement in crime is shown to be related inversely to deterrence variables and other opportunity costs associated with crime and directly to the differential return it can provide over legitimate activity as well as to risk aversion. Contrary to the perception that criminals constitute a "noncompeting group" in classical labor jargon, by which all of them are completely specialized in the pursuit of criminal activities as members of gangs or organized crime, the labor-theoretic approach to participation in illegitimate activities expects many offenders to be "part-time" offenders, pursuing legitimate endeavors as well, and criminal enterprises to be typically small organizations, partly to diversify excessive risk bearing, given the prospect of detection, apprehension, and punishment (see Ehrlich, 1973, and the extensive empirical evidence documented by Reuter, MacCoun, & Murphy, 1990).

The theory yields not just general qualitative propositions about the way offenders respond to incentives but also discriminating implications about the relative magnitudes of responses to different incentives. For example, a 1% increase in the probability of apprehension is shown to exert a larger deterrent effect than corresponding increases in the conditional probabilities of conviction and punishment. Essentially due to conflicting income and substitution effects, however, sanction severity can have more ambiguous effects on active offenders: A strong preference for risk may weaken or even reverse the deterrent effect of sanctions, and the results are even less conclusive if one assumes that the length of time spent in crime, not just the moral obstacle to entering it, generates disutility. The results become less ambiguous at the aggregate level, however, as one allows for heterogeneity of offenders due to differences in employment opportunities or preferences for risk and crime: A more severe sanction can reduce the crime rate by deterring the entry of potential offenders even if it has little effect on actual ones. In addition to heterogeneity across individuals in personal opportunities and preferences, the literature has also addressed the role of heterogeneity in individuals' perceptions about probabilities of apprehension, as affected by learning from past experience. As a result, current crime rates may react, in part, to past deterrence measures. A different type of heterogeneity that can affect variations in crime across different crime categories and geographical units may stem from the degree of social interaction, which can partly explain why urban crime rates generally exceed rural rates.

This theory also yields testable propositions about the way participation in criminal activity varies across states,

over the life cycle, and across different crime categories. Ehrlich's work identifies inequality in the distribution of income, especially at the lower tail of the income distribution, as having a powerful impact on participation in all felonies, essentially because those with lower skills have poorer prospects for entry into legitimate occupations and lower opportunity costs of imprisonment. In contrast, area wealth provides higher gains from property crimes. Also, educated workers who earn relatively high salaries in legitimate occupations can be expected to avoid especially street crimes, which require low skills, but this is not necessarily the case with white-collar crimes, which do require education and legitimate skills. Evidence from prison data and self-reported crimes confirms these propositions (Ehrlich, 1973, 1975a; Lochner, 2004).

## Private "Demand"

The incentives operating on offenders often originate from, and are partially controlled by, consumers and potential victims. Transactions in illicit drugs or stolen goods, for example, are patronized by consumers who generate a direct demand for the underlying offense. But even crimes that inflict pure harm on victims are affected by an indirect (negative) demand, or "derived demand," which is derived from a positive demand for safety, or "self-protection." This term has been used in the economic literature to indicate individual efforts aimed at reducing the probability of being afflicted by hazards to their economic or physical well-being (Ehrlich & Becker, 1972). The hazard of becoming a victim to crime is a natural application of the concept. By their choice of optimal self-protective efforts through use of locks, safes, Lojacks, private guards and alarm systems, or selective avoidance of crime-prone areas, potential victims lower the marginal returns to offenders from targeting them and thus the implicit return on crime to the offender. And since optimal self-protection generally increases with the perceived risk of victimization (the crime rate), private protection and public enforcement will be interdependent. Thus, even in the absence of public enforcement of laws, the incidence of crime in the population can be contained or "equilibrated" through private self-protection. At the same time, however, private protection can also generate both positive and negative externalities (i.e., spillover effects) of considerable magnitudes, which makes it both necessary and expedient to resort to the power of the state to play the major role in protecting life and property and ensuring law and order.

## Public Intervention

Since crime, by definition, causes a net social loss, and crime control measures are largely a public good, collective action is needed to augment individual self-protection. Public intervention typically aims to "tax" illegal returns through the threat of punishment or to "regulate" offenders, via incapacitation and rehabilitation programs. All control measures are costly. Therefore, the "optimum" volume of offenses cannot be nil but must be set at a level where the marginal cost of each measure of enforcement or prevention equals its marginal benefit.

To assess the relevant net social loss, however, one must adopt a criterion for public choice. Becker (1968) and Stigler (1970) have chosen maximization of variants of "social income" measures as the relevant criterion, requiring the minimization of the sum of social damages from offenses and the social cost of law enforcement activities. This approach leads to powerful propositions regarding the optimal magnitudes of probability and severity of punishments for different crimes and different offenders or, alternatively, the optimal level and mix of expenditures on police, courts, and corrections. The analysis also reaffirms the classical utilitarian proposition that the optimal severity of punishment should "fit the crime" and thus be set according to its overall deterrent value, essentially because applying the more severe sanctions on, say, petty theft would induce offenders to go for grand larceny. Moreover, it makes a strong case for the desirability of monetary fines, when feasible, as a deterring sanction: Since fines are essentially a transfer payment that does not require use of real resources to punish offenders as do imprisonment, confinement, deportation, and banishments, they are "socially costless." However, fines cannot be relied upon as the dominating form of sanctions since optimal crime control requires the reliance on incapacitating offenders with a high risk of recidivism or providing them opportunities for rehabilitation through training programs oriented to bolster legitimate skills (Ehrlich, 1981).

Another aspect of optimal law enforcement where the income-maximizing, or cost-minimizing, criterion has a natural appeal is the choice of the optimal way to produce enforcement or security services. This is inherently a "supply-side" issue since optimal enforcement services need not be produced by government agencies—they could be delivered by private enforcers or private suppliers of security and protection services whose task can be to detect legal infractions, help prosecute offenders, or administer legal sanctions. Optimal public law enforcement involves setting up rules of compensation to maximize the efficiency of the supply of enforcement and protection services, including private provision of detention, imprisonment, and rehabilitation services (see, e.g., Becker & Stigler, 1974; Benson, 1998; Landes & Posner, 1975).

Different criteria for public choice, however, yield different implications regarding the optimal mix of law enforcement strategies. An important example is the optimal mix of probability and severity of apprehending and punishing offenders, as is the case when the social welfare function is expanded to include concerns for "distributional consequences" of law enforcement on offenders and victims in addition to aggregate income. These considerations can be ascribed to aversion to risk, as in Polinsky and

Shavell (1979), or to an alternative concept of justice, as proposed in Ehrlich (1982). Furthermore, a positive analysis of enforcement must address the behavior of the separate agencies constituting the enforcement system: police, courts, and prison authorities. For example, Landes's (1971) analysis of the courts, which focuses on the interplay between prosecutors and defense teams, explains why settling cases out of court may be an efficient outcome of many court proceedings.

The optimal enforcement policy arising from the income-maximizing criterion can be questioned from yet another angle: a public-choice perspective. The optimization rule invoked in the preceding papers assumes that enforcement is carried out by a social planner. In practice, public law enforcement can facilitate the interests of rent-seeking enforcers who are amenable to malfeasance and bribes. Optimal social policy needs to control malfeasance by properly remunerating public enforcers (Becker & Stigler, 1974) or setting, where appropriate, milder penalties (Friedman, 1999).

### Market Equilibrium

In Ehrlich's (1981) "market model," the equilibrium flow of offenses results from the interaction between aggregate supply of offenses, direct or derived demand for offenses (through self-protection), and optimal public enforcement, which operates like a tax on criminal activity. Some behavior classified as crime, such as prostitution and consumption of illicit drugs, involves the interaction between suppliers and consumers in an explicit market setting. But even crimes against persons and property can be analyzed by reference to a virtual market that involves the interaction between offenders and potential victims through the latter's demand for self-protection and thus a negative demand for crime. This analysis identifies more fully the interaction between crime, private self-protection, and public law enforcement and the limitations of alternative means of crime prevention. One important application concerns a comparison of deterrence, incapacitation, and rehabilitation as instruments of crime control. This is because the efficacy of deterring sanctions cannot be assessed merely by the elasticity of the aggregate supply of offenses schedule, as it depends on the elasticity of the private demand schedule as well. Likewise, the efficacy of rehabilitation and incapacitation programs cannot be inferred solely from knowledge of their impact on individual offenders. It depends crucially on the elasticities of the market supply and demand schedules, as these determine the extent to which successfully rehabilitated offenders will be replaced by others responding to the prospect of higher net returns. This market setting has also been applied in works by Schelling (1967), Buchanan (1973), and Garoupa (2000), for example, to analyze various aspects of organized crime.

The "market model" has been developed largely in a static, partial-equilibrium setting in which the general economy affects the illegal sector of the economy but not vice versa. More recently, the model has been extended to deal with the interaction between the two under dynamic settings as well. Specific applications focus on the interaction between crime and income distribution and the relation between bureaucratic corruption and economic growth over the process of economic development (Ehrlich & Lui, 1999; Imrohoroglu, Merlo, & Rupert, 2000).

## Applications and Empirical Evidence

### Scope

Largely due to the paucity of theoretically relevant data, little has been done thus far to implement a comprehensive market model of illegitimate activity. Many researchers have attempted, however, to implement a simultaneous equation model of crime and law enforcement activity consisting, typically, of three sets of structural equations originally proposed in Ehrlich (1973): supply-of-offenses functions linking the rate of offenses with deterrence variables and other measurable incentives; production functions of law enforcement activity linking conditional probabilities of arrest, conviction, and punishment with resource inputs and other productivity measures; and demand-for-enforcement functions linking resource spending with determinants of public intervention. Attempts to address other aspects of the market model of crime include measuring the effect of law enforcement on the net return from crime, as in Viscusi (1986), and simulating a general equilibrium model that focuses on the interaction between the legal and illegal sectors of the economy (Engelhardt, Rocheteau, & Rupert, 2008; Imrohoroglu et al., 2000). Other applications concern modeling organized crime or "victimless crimes," for which there is direct demand by consumers and patrons, and analyses of the general criminal justice system (see, e.g., Buchanan, 1973; Garoupa, 2000; Reinganum, 1993).

A number of studies have focused instead on specific components of the market model. Some have examined various forms of private self-protection, including use of guards, flight from high-crime neighborhoods into suburbs, and carrying concealed guns as a possible means of self-protection against crime. Other studies have focused on the effects of specific measures of law enforcement, such as the federal sentencing guidelines, enhanced police presence, and the two- and three-strike legislation in California. Another set of papers focuses on crime and various aspects of the labor market. Studies related to these applications by Bartel (1975, 1979); Ayres and Levitt (1998); Cullen and Levitt (1999); Lott and Mustard (1997); Duggan (2001); LaCass and Payne (1999); Gould, Weinberg, and Mustard (2002); Raphael and Winter-Ebmer (2001); Shepherd (2002); and Lochner (2004) are surveyed in Ehrlich and Liu (2006, vols. 2 and 3). See also Kling (2006) and Evans and Owens (2007).

The bulk of the empirical implementation, especially over the past two decades, tested the deterrence hypothesis against data from different population aggregates and different types of crime. The first set explores data on individual offenders, juveniles, females, urban areas, and different countries, such as Canada, the United Kingdom, Italy, Finland, and Germany. The second set includes data on specific crime categories, ranging from aircraft hijacking, drugs, drunk driving, antitrust violations, corporate fraud, and federal fraud. Early and recent papers on these topics (e.g., Bartel, 1975, 1979; Block, Nold, & Sidak, 1981; Corman & Mocan, 2000; Glaeser & Sacerdote, 1999; Karpoff & Lott, 1993; Landes, 1971; Levitt, 1997; Waldfogel, 1995; Witte, 1980; Wolpin, 1978) are also surveyed in Ehrlich and Liu (2006, vols. 2 and 3). Also see the recent study by Drago, Galbiati, and Vertova (2009).

## Methodological Issues and Major Findings

The econometric applications via regression analyses have been hampered by a number of methodological problems. Federal Bureau of Investigation (FBI) crime reports are known to understate true crime rates, and related errors of measurement in estimated punishment risks may expose parameter estimates to biases and spurious correlations. The inherent simultaneity in the data, whereby the relation between crime and deterrence variables may reflect different directions of causality corresponding to supply, demand, or production relationships, requires systematic use of identification restrictions to ensure consistent estimation of structural parameters. In testing offenders' responsiveness to incentives, the deterrent effect of imprisonment must be distinguished from its incapacitating effect. Efficient functional forms of structural equations must be selected systematically. And there is the ubiquitous possibility that regression estimates would be affected by "missing variables" and sample selection biases. Most of these problems have been recognized and addressed in the literature from the outset, but more attention has been paid to these problems in recent studies.

In particular, and as the "market model" of crime suggests, higher crime rates, and thus risks of victimization, should increase the willingness of potential victims and law enforcement agencies to spend resources on crime prevention and deterrence, which could bias the estimated association between crime and enforcement variables in a direction contrary to the one predicted by the deterrence hypothesis. In contrast, the "crowding effect" on existing law enforcement resources that can be produced by unexpected surges of crime, as well as errors of measurement in crime counts, are likely to produce the opposite bias (Ehrlich, 1973). To overcome such statistical biases, researchers must try to identify instrumental variables that are not affected by the concurrent incidence of crime but otherwise raise the public willingness to enforce the law or the efficiency of law enforcement efforts. Examples of variables identified by researchers as "instrumental variables" include "political cycles" that bring to power politicians running on strengthening law and order (see, e.g., Levitt, 1997, although successful campaigns may also reflect previously high and persisting crime rates) and the stock of legal decisions stemming from constitutional principles, especially by the Supreme Court, which make it either easier or more difficult to convict those apprehended and charged by police (see, e.g., Ehrlich & Brower, 1987). Another important research challenge concerns the separation of deterrence from incapacitation effects since law enforcement, by way of detention, incarceration, and imprisonment, can reduce the incidence of crime by incapacitating actual offenders rather than deterring would-be ones. The deterrence hypothesis applies only to the latter effect. The incapacitation effect can be assessed theoretically, since it depends largely on the level of the risks of apprehension and imprisonment rather than changes in these variables (Ehrlich, 1982), but the two effects can also be isolated empirically by estimating a set of interrelated crimes simultaneously: For example, higher prison terms for burglary cannot have incapacitating effects on those committing robbery, but they may deter would-be robbers (Kessler & Levitt, 1999). Helland and Tabarrok (2007) provide further evidence on the existence of pure deterrence effects based on California's "three-strike" law.

Use of different types of data can also affect the researcher's ability to measure the effects of "positive incentives." For example, cross-sectional data on variations in unemployment rates may be affected by an area's industrial composition, unemployment compensation level, and the age and skill composition of the area's labor force rather than involuntary layoffs. Time-series data that span business cycles are more likely to reflect involuntary layoffs and thus the opportunities for gains from legitimate labor market activities. Studies based on time-series and panel data are therefore more likely to reflect the force of legal employment incentives—the effect of the unemployment rate specifically—and the deterrence hypothesis.

The overwhelming volume of studies following systematic econometric applications, which were applied to alternative regions, population groups, and different crime categories, has produced similar findings: Probability and length of punishment are generally found to lower crime rates, with elasticities of response of crime rates to probability of punishment often exceeding those with respect to severity of punishment (see, e.g., the early studies by Ehrlich and the recent study by Drago et al., 2009). Also, the estimated elasticities of crime with respect to the risk of apprehension are generally found to exceed those with respect to the conditional risks of conviction and punishment. Crime rates are also found to be directly related to measures of income inequality and community wealth (proxy measures of relative gains from crime). Estimates of unemployment effects are somewhat ambiguous, however, depending, in part, on whether they are derived from

time-series or cross-section data (see the surveys by Ehrlich, 1996; Freeman, 1983), but more recent studies (e.g., Raphael & Winter-Ebmer, 2001) have confirmed the existence of deterrent effects of employment opportunities using instrumental variable techniques.

Not all past research appears to be consistent with the deterrence hypothesis. For example, some studies report a positive association between police expenditure and crime, although this relationship may represent positive demand for public protection from crime. Critics have argued that the estimated deterrent effects may mask a crowding effect of crime on punishment rather than vice versa (Blumstein, Cohen, & Nagin, 1978). However, recent studies provide evidence corroborating the deterrence hypothesis by using effective instrumental variables and panel data techniques to isolate and identify the effect of deterrence variables. Some studies have also estimated a system of interrelated crimes to separate deterrence from incapacitation effects of imprisonment (e.g., Ehrlich & Brower, 1987; Ehrlich & Liu, 1999; Kessler & Levitt, 1999; Levitt, 1997).

The applicability of the economic approach to the crime of murder, and whether the death penalty constitutes a specific deterrent have raised greater controversy. The center of debate has been Ehrlich's (1975a, 1977) studies based on time-series and cross-state data, in which deterrence variables, including the risk of execution as well as the probability and length of punishment by imprisonment, were found to lower murder rates (see Blumstein et al., 1978; and the response in Ehrlich & Mark, 1977). The controversy has generated additional empirical research, which is still ongoing, some inconsistent with the deterrence hypothesis (e.g., Avio, 1979; Hoenack & Weiler, 1980; McManus, 1985) but others strongly corroborative of not only the direction of the deterrent effect of the probability of execution but even the quantitative impact of this extreme penalty as originally estimated (e.g., Dezhbakhsh, Rubin, & Shepherd, 2003; Ehrlich & Liu, 1999; Layson, 1983, 1985; Mocan & Gittings, 2003; Wolpin, 1978). The studies by most economists have been motivated by scientific curiosity concerning the issue of deterrence, for if severity of punishment matters, as is overwhelmingly documented by studies following a rigorous econometric methodology, why wouldn't the most severe legal sanction impart a deterrent effect as well? Many critiques have been motivated, however, by a normative concern about the desirability of executions as a legal sanction, an issue that should be divorced from the issue of its effectiveness. Indeed, if murderers respond to incentives, alternative sanctions can be effective as well.

## Policy Implications

The principle of minimizing the aggregate income loss from crime, while involving a narrow efficiency criterion involving material losses, nevertheless has powerful implications concerning optimal crime control policies. It suggests the need to employ differential punishments for different types of offenses, to "fit the crime," as well as for different types of offenders in proportion to their deterrent and incapacitating or rehabilitating values. It also implies that although government intervention in the economy is necessitated to avoid a suboptimal reliance on private self-protection, which creates both external economies (thus too little private protection) and diseconomies (too much emphasis on the protection of individual interests as opposed to society as a whole), the production of means of crime control need not be done just by government agencies and may involve outsourcing the production of desirable security services to private firms (e.g., bail bonding services, legal firms, private prisons, and training centers).

The relative desirability of specific means of crime control cannot be determined just by their relative efficacy or cost-efficiency; it also depends on their relative social costs and on the welfare criteria invoked as a justification for public law enforcement. For example, if the welfare objective is to maximize social income, then the social cost of purely deterring sanctions, such as fines, would be close to zero because as transfer payments, fines are free of the deadweight losses associated with imprisonment, house arrests, probation, and other intermediate punishments. An optimal enforcement strategy may then involve raising such fines to their maximal feasible level (consistent with a convict's wealth constraint) while lowering the probability of apprehension and conviction to its minimal level. Even under this (narrow) efficiency criterion, however, it would be optimal to use imprisonment and intermediate punishments along with fines for those crime categories where the added incapacitation value of imprisonment justifies its added costs.

The enforcement strategy would be different if the social welfare function were broadened to include distributional objectives as well. These include, for example, a preference for promoting equality of individuals under the law, reducing the legal error of convicting the innocent, or lowering the corollary prospect of letting the guilty go free. For example, since the probability of apprehension and punishment is substantially less than 1, penalties are in fact applied through a lottery system. Offenders who are caught and punished are subjected to ex post discrimination under the law because they "pay" not just for their own crime but also for offenders who get away with crime. The degree of such ex post discrimination rises as the penalty becomes more severe or if the probability of punishment is very low. Such concerns help explain why severity of punishment is often traded on the margin for a higher probability of apprehension and conviction. It also helps explain why the justice system introduces numerous safeguards to protect the rights of the accused and why the opposition to capital punishment tends to increase when the penalty is applied infrequently and capriciously (Ehrlich, 1982).

Incorporating concerns for equality and legal error in the social welfare function raises not just the marginal social cost of severity of punishment but that of any strategy of enforcement (as long as the probability of being arrested and punished for a crime is low) relative to its cost under the narrower efficiency criterion. The implication is that more crime would be tolerated as a result of a trade-off between equity and efficiency in enforcement—a trade-off typical of social choice in general.

Another common mistake is the argument that higher severity of punishment makes juries less inclined to convict or punish severely or that more public law enforcement will necessarily lower private protection. There is, however, no mechanical substitution relationship between severity of penalty and probability of conviction and punishment, and the association, or simple correlation, between these two variables could be both positive and negative since it depends on the underlying factors that generate their observed co-variation. If the more severe punishment is considered in cases where optimal severity should be lower, as is the case when the crime is less severe or the penalty is less likely to deter because of the offender's age or mental capacity, then indeed, juries would be less inclined to convict. However, if severity of punishment is justified by the greater severity of the offense or a higher risk of victimization in the community, the more severe penalty could be associated with a greater tendency to convict (see Ehrlich & Gibbons, 1977).

This analysis is applicable to crime control strategies concerning the use of positive incentives as well. The market model implies that a lower disparity in the distribution of earning opportunities in legitimate markets will deter offenders on the margin by reducing their differential gains from criminal activity. This provides a justification for public policies aimed at equalizing educational and employment opportunities partly as means of reducing crime. However, since these policies, unlike conventional law enforcement, cannot be targeted specifically at actual or potential offenders, they may entail relatively high social costs as means of crime control. The positive implications of the market model and some corroborating empirical evidence concerning the relative efficacy of deterrence versus incapacitation and rehabilitation for many crimes suggest a direction of reform of the criminal justice system through greater reliance on general incentives and purely deterring sanctions. Forcing offenders to pay fines through work release programs (including direct restitution to their victims) may in many cases be as effective a means of crime prevention as the more costly incapacitating penalties—especially in the case of many theft crimes or transactions in illicit goods and services. The dramatic growth in the proportion of those imprisoned for drug offenses in the past few decades appears to be inconsistent with this implication of optimal enforcement.

"Intermediate punishments," including work release programs and probation, can also provide effective substitutes to imprisonment and incapacitation through the force of incentives, as implied by the general "deterrence hypothesis."

## Future Directions

The application of the economic approach to crime in the analysis of some of the main components of the criminal justice system has been done primarily under a partial equilibrium setting in which the criminal sector of the economy has been analyzed separately from the general economy. The analysis has revolutionized previous literature on the causes and consequences of criminal activity by identifying new factors that account for the diversity of the incidence of criminal behavior across geographical areas, across population groups, and over time, as well as by predicting and measuring their empirical importance. Much remains to be done, however, to improve both the analytical rigor and econometric accuracy of the relevant behavioral relations and clarify their implications for public policy. On the analytical part, the complete application of the market model, which exposes more fully the interaction of crime with both private and public protection efforts, is still a challenge to be addressed, partly through better collection and utilization of the relevant data. Moreover, the development of a general equilibrium setting that identifies more fully the interaction between the legal and illegal sectors of the economy and their feedback effects at a point in time, over the business cycle, and over the long run are still challenges to be met. Indeed, macroeconomic and growth accounting still misses a satisfactory assessment of the role played by the illegal sector of the economy, the underground economy in particular, on the behavior of the full economy and its dynamic movements. I expect the next wave of works on crime to delve into these challenging areas of inquiry through the application of general equilibrium models and numerical analyses under both static and dynamic settings.

Also, a disproportionate work on illegal activity has focused so far on felonies and street crime and less on white-collar crimes because of the paucity of relevant statistical data on the latter: At this point, there is little information available on the actual volume of these legal infractions independently of clearance and arrest statistics. The relatively little attention paid so far to fraud, digital counterfeiting, and violations of fiduciary obligations in business endeavors may have been partly the result of unjustified belief in the power of competitive market forces to eliminate these economically and socially harmful behaviors, but the recent worldwide financial crisis implies that the study of these illegal infractions deserves closer attention. The greater reliance in recent decades on victimization studies and comprehensive population surveys could offer opportunities for systematic applications of the economic approach in studying white-collar crimes and their economic impact.

Furthermore, most of the studies on crime have relied on within-country data. The application of the economic approach in studying variation in criminal activity across countries and different legal systems is still a major challenge to be met, largely because of differences in the definition, interpretations, and methods of reporting crime. The role of the legal system itself (e.g., the common law, the Islamic Sharia, or the Napoleonic Code) is yet to be revealed through rigorous analyses.

## Conclusion

The economic approach to crime has provided one of the important revolutions in the social sciences by applying the rigorous tools of economic analysis and econometric methodology to offer a unified approach for understanding illegal behavior as part of human behavior in general. It has also offered significant insights about the relative efficiency and desirability of different means of crime control and different components of the law enforcement system.

The economic approach and the "market model" specifically, however, are still a work in progress. It is still too early to assess the degree to which the various econometric studies have produced accurate estimates of critical behavioral relationships underlying variations and trends of crime. While a strong consensus is emerging in the economic literature regarding the power of the economic approach to explain criminal behavior and the validity of the "deterrence hypothesis," future progress will depend on better data and more complete implementations of the comprehensive model of crime.

The economic approach, however, has not been equally embraced outside of economics. This is partly due to healthy interdisciplinary competition. Criminologists have also tended to produce theories of crime in which the economic and social environments and institutions, rather than individual incentives and enforcement efforts, play the major role in explaining the phenomenon. This, however, does not contradict the economic approach, which assigns an important role to institutions as well, often as an endogenously evolving part of the comprehensive market system. A common misconception about the broad meaning of the "deterrence hypothesis" is that it applies only to negative incentives, while positive incentives may hold a greater promise for solving the crime problem. Another often heard claim is that we don't need to know more about punishment because punishment does not eliminate crime. Both claims are inappropriate. The deterrence hypothesis and its logical extension—the market model—rely on the marginal efficacy of both positive and negative incentives and on the interaction between market demand and supply forces to explain the observed variability in the frequency of offenses across space and time. The empirical evidence developed in most econometric applications is consistent with the hypothesis that punishment and other general incentives exert a deterrent effect on offenders. This suggests, for example, that there is no need to rely exclusively on harsh or incapacitating sanctions to achieve efficient crime control. A better understanding of what does work, however, calls for more rather than less research into the general deterrence hypothesis and the market model based on it.

## References and Further Readings


Avio, K. L. (1979). Capital punishment in Canada: A time-series analysis of the deterrent hypothesis. *Canadian Journal of Economics, 12,* 647–676.

Ayres, I., & Levitt, S. D. (1998). Measuring positive externalities from unobservable victim precaution: An empirical analysis of Lojack. *Quarterly Journal of Economics, 113,* 43–77.

Bartel, A. P. (1975). An analysis of firm demand for protection against crime. *Journal of Legal Studies, 4,* 433–478.

Bartel, A. P. (1979). Women and crime: An economic analysis. *Economic Inquiry, 17,* 29–51.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy, 76,* 169–217.

Becker, G. S., & Landes, W. M. (Eds.). (1974). *Essays in the economics of crime and punishment.* New York: Columbia University Press.

Becker, G. S., & Stigler, G. J. (1974). Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies, 3,* 1–18.

Benson, B. (1998). *To serve and protect: Privatization and community in criminal justice.* New York: New York University Press.

Block, M. K., & Heineke, J. M. (1975). A labor theoretic analysis of the criminal choice. *American Economic Review, 65,* 314–325.

Block, M. K., Nold, F. C., & Sidak, J. G. (1981). The deterrent effect of antitrust enforcement. *Journal of Political Economy, 89,* 429–445.

Blumstein, A., Cohen, J., & Nagin, D. (Eds.). (1978). *Deterrence and incapacitation: Estimating the effects of criminal sanctions on crime rates.* Washington, DC: National Academy of Science.

Buchanan, J. M. (1973). A defense of organized crime? In *The economics of crime and punishment: A conference sponsored by American Enterprise Institute for Public Policy Research* (pp. 119–132). Washington, DC: American Enterprise Institute for Public Policy Research.

Clotfelter, C. T. (1977). Public services, private substitutes, and the demand for protection against crime. *American Economic Review, 67,* 867–877.

Cook, P. J. (1975). The correctional carrot: Better jobs for parolees. *Policy Analysis, 1,* 11–55.

Corman, H., & Mocan, H. N. (2000). A time-series analysis of crime, deterrence, and drug abuse in New York City. *American Economic Review, 90,* 584–604.

Cullen, J. B., & Levitt, S. D. (1999). Crime, urban flight, and the consequences for cities. *Review of Economics and Statistics, 81,* 159–169.

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,* 344–376.

Drago, F., Galbiati, R., & Vertova, P. (2009). The deterrent effects of prison: Evidence from a natural experiment. *Journal of Political Economy, 117,* 257–280.

Duggan, M. (2001). More guns, more crime. *Journal of Political Economy, 109,* 1086–1114.

Ehrlich, I. (1973). Participation in illegitimate activities: Theoretical and empirical investigation. *Journal of Political Economy, 81,* 521–565. Reprinted with supplements in Becker and Landes (1974).

Ehrlich, I. (1975a). The deterrent effect of capital punishment: A question of life and death. *American Economic Review, 65,* 397–417.

Ehrlich, I. (1975b). On the relation between education and crime. In F. T. Juster (Ed.), *Education, income, and human behavior*. Cambridge, MA: National Bureau of Economic Research.

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,* 307–322.

Ehrlich, I. (1982). The optimum enforcement of laws and the concept of justice: A positive analysis. *International Review of Law and Economics, 2,* 3–27.

Ehrlich, I. (1996). Crime, punishment, and the market for offenses. *Journal of Economic Perspectives, 10,* 43–67.

Ehrlich, I., & Becker, G. S. (1972). Market insurance, self-insurance, and self-protection. *Journal of Political Economy, 80,* 623–648.

Ehrlich, I., & Brower, G. D. (1987). On the issue of causality in the economic model of crime and law enforcement: Some theoretical considerations and experimental evidence. *American Economic Review, Papers and Proceedings, 77,* 99–106.

Ehrlich, I., & Gibbons, J. (1977). On the measurement of the deterrent effect of capital punishment and the theory of deterrence. *Journal of Legal Studies, 6,* 35–50.

Ehrlich, I., & Lui, F. T. (1999). Bureaucratic corruption and endogenous economic growth. *Journal of Political Economy, 107*(Pt. 2), S270–S293.

Ehrlich, I., & Liu, Z. (1999). Sensitivity analyses of the deterrence hypothesis: Let's keep the econ in econometrics. *Journal of Law and Economics, 42*(Pt. 2), 455–487.

Ehrlich, I., & Liu, Z. (2006). *The economics of crime* (3 vols.). Cheltenham, UK: Edward Elgar.

Ehrlich, I., & Mark, M. R. (1977). Fear of deterrence. *Journal of Legal Studies, 6,* 293–316.

Engelhardt, B., Rocheteau, G., & Rupert, P. (2008). Crime and the labor market: A search model with optimal contracts. *Journal of Public Economics, 92,* 1876–1891.

Entorf, H., & Spengler, H. (2000). Socioeconomic and demographic factors of crime in Germany: Evidence from panel data of the German States. *International Review of Law and Economics, 20,* 75–106.

Evans, N., & Owens, E. (2007). Cops and crime, *Journal of Public Economics, 91,* 181–201.

Forst, B. E. (1976). Participation in illegitimate activities: Further empirical findings. *Policy Analysis, 2,* 477–492.

Freeman, R. B. (1983). Crime and unemployment. In J. Q. Wilson (Ed.), *Crime and public policy* (pp. 89–106). San Francisco: Institute for Contemporary Studies.

Freeman, R. B. (1996). Why do so many young American men commit crimes and what might we do about it? *Journal of Economic Perspectives, 10,* 25–42.

Friedman, D. (1999). Why not hang them all: The virtues of inefficient punishment. *Journal of Political Economy, 107* (Pt. 2), S259–S269.

Garoupa, N. (2000). The economics of organized crime and optimal law enforcement. *Economic Inquiry, 38,* 278–288.

Glaeser, E. L., & Sacerdote, B. (1999). Why is there more crime in cities? *Journal of Political Economy, 107*(Pt. 2), S225–S258.

Gould, E. D., Weinberg, B. A., & Mustard, D. B. (2002). Crime rates and local labor market opportunities in the United States: 1979–1997. *Review of Economics and Statistics, 84,* 45–61.

Grogger, J., & Willis, M. (2000). The emergence of crack cocaine and the rise in urban crime rates. *Review of Economics and Statistics, 82,* 519–529.

Hannan, T. H. (1982). Bank robberies and bank security precautions. *Journal of Legal Studies, 11,* 83–92.

Helland, E., & Tabarrok, A. (2007). Does three strikes deter? A non-parametric investigation. *Journal of Human Resources, 42,* 309–330.

Hoenack, S. A., & Weiler, W. C. (1980). A structural model of murder behavior. *American Economic Review, 70,* 327–341.

Imrohoroglu, A., Merlo, A., & Rupert, P. (2000). On the political economy of income redistribution and crime. *International Economic Review, 41,* 1–25.

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.

Kessler, D., & Levitt, S. D. (1999). Using sentence enhancements to distinguish between deterrence and incapacitation. *Journal of Law and Economics, 42*(Pt. 2), 343–363.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review, 96,* 863–876.

LaCass, 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*(Pt. 2), 245–269.

Landes, W. M. (1971). An economic analysis of the courts. *Journal of Law and Economics, 14,* 61–107.

Landes, W. M., & Posner, R. A. (1975). The private enforcement of law. *Journal of Legal Studies, 4,* 1–46.

Layson, S. (1983). Homicide and deterrence: Another view of the Canadian time-series evidence. *Canadian Journal of Economics, 16,* 52–73.

Layson, S. (1985). Homicide and deterrence: A reexamination of the United States time-series evidence. *Southern Journal of Economics, 52,* 68–89.

Levitt, S. D. (1997). Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review, 87,* 270–290.

Liu, Z. (2004). Capital punishment and the deterrence hypothesis: Some new insights and empirical evidence. *Eastern Economic Journal, 30,* 237–258.

Lochner, L. (2004). Education, work, and crime: A human capital approach. *International Economic Review, 45,* 811–843.

Lott, J. R., & Mustard, D. B. (1997). Crime, deterrence, and right-to-carry concealed handguns. *Journal of Legal Studies, 26,* 1–68.

McManus, W. S. (1985). Estimates of the deterrent effect of capital punishment: The importance of the researcher's prior beliefs. *Journal of Political Economy, 93,* 417–425.

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, 42,* 453–478.

Philipson, T. J., & Posner, R. A. (1996). The economic epidemiology of crime. *Journal of Law and Economics, 39,* 405–433.

Phillips, L. (1981). The criminal justice system: Its technology and inefficiencies. *Journal of Legal Studies, 10,* 363–380.

Polinsky, A. M., & Shavell, S. (1979). The optimal trade-off between the probability and magnitude of fines. *American Economic Review, 69,* 880–891.

Raphael, S., & Winter-Ebmer, R. (2001). Identifying the effect of unemployment on crime. *Journal of Law and Economics, 44,* 259–283.

Reinganum, J. F. (1993). The law enforcement process and criminal choice. *International Review of Law and Economics, 13,* 115–134.

Reuter, P., MacCoun, R., & Murphy, P. (1990). Money from crime: A study of the economics of drug dealing in Washington, D.C. *Contemporary Drug Problems, 18,* 713–716.

Sah, R. K. (1991). Social osmosis and patterns of crime. *Journal of Political Economy, 99,* 1272–1295.

Schelling, T. C. (1967). Analysis of organized crime. In *Task force report: Organized crime. The President's Commission on Law Enforcement and Administration of Justice* (pp. 114–126). Washington, DC: Government Printing Office.

Shepherd, J. M. (2002). Fear of the first strike: The full deterrent effect of California's two- and three-strikes legislation. *Journal of Legal Studies, 31*(Pt. I), 159–201.

Stigler, G. J. (1970). The optimum enforcement of laws. *Journal of Political Economy, 78,* 526–535.

Van den Haag, E. (1975). *Punishing criminals.* New York: Basic Books.

Viscusi, W. K. (1986). The risks and rewards of criminal activity: A comprehensive test of criminal deterrence. *Journal of Labor Economics, 4*(Pt. I), 317–340.

Wadycki, W. J., & Balkin, S. (1979). Participation in illegitimate activities: Forst's model revisited. *Journal of Behavioral Economics, 8,* 151–163.

Wahlroos, B. (1981). On Finnish property criminality: An empirical analysis of the postwar era using an Ehrlich model. *Scandinavian Journal of Economics, 83,* 553–562.

Waldfogel, J. (1995). Are fines and prison terms used efficiently? Evidence on federal fraud offenders. *Journal of Law and Economics, 38,* 107–139.

Witte, A. D. (1980). Estimating the economic model of crime with individual data. *Quarterly Journal of Economics, 94,* 57–84.

Wolpin, K. I. (1978). Capital punishment and homicide in England: A summary of results. *American Economic Review, Papers and Proceedings, 68,* 422–427.

# 74

# Economics of Property Law

Richard D. Coe
*New College of Florida*

Property law is the body of law that establishes the rules governing ownership rights over scarce resources. Ownership rights are multifaceted—often referred to as a "bundle of rights"—but in general such rights encompass three broad areas of control over a specific resource: the right to exclude (the ability of the owner to prevent others from using the resource), the right to use (the ability of the owner to use the resource in the manner he or she sees fit), and the right to transfer (the ability of the owner to assign ownership rights to the resource to another). The economic analysis of property law is primarily concerned with the effect of various property rules on the allocation of resources and whether such effects conform to the economic concept of efficiency.[1] Equity issues are also addressed but to a lesser degree.

## The Economics of the Right to Exclude

Ownership rights to a resource grant to the owner the ability to exclude others from using the resource. In contrast, an open-access resource, often labeled common property, is one that an individual has a right to use but cannot exclude others from using. The ability to exclude is probably the most commonly associated aspect of ownership—"that's mine; you can't use it." A derivative power of the excludability right is that the owner can determine which other persons can use the property, either now or in the future, via some form of permission, such as a rental or licensing agreement.

### The Minimization of Conflict Costs

Economists distinguish between goods and resources that are characterized by rival consumption (or use) and goods and resources that are characterized by nonrival consumption (or use). Rival consumption exists when consumption by one individual physically precludes consumption by another individual, such as the eating of an apple. Nonrival consumption exists when consumption by one person does not preclude consumption by others, such as the viewing of a fireworks display.

A world of scarce resources characterized by rival use will inevitably lead to disputes over the control of such resources. If ownership rights are uncertain, costly conflicts arise in an attempt to gain or defend use of the resource. Resources are expended (including lives lost) in the actual conflict as well as in preemptive protective measures against attack. Conflict costs are particularly detrimental in that they are unproductive from a societal standpoint. Nothing new is produced; worse yet, existing valuable resources are destroyed—a negative-sum game. Thomas Hobbes (1651/1963) was one of the earliest political philosophers to emphasize the costs of a lawless "state of nature," where ownership rights were determined by the private use of force. A well-defined system of ownership rules can reduce the uncertainty and attendant conflict costs regarding who controls scarce resources. While it requires resources to establish and enforce a system of property rights, a state-enforced system has substantial economies of scale over a system where each individual is responsible for establishing and defending his or her ownership rights.

In contemporary society, most ownership rights are acquired via transfer from the (previous) legal owner of the resource. However, property law is regularly called upon to resolve ownership rights in resources that have no prior owner. *Pierson v. Post* (1805) is a legendary case on this issue. Post was leading a fox hunt. As the target was in sight, Pierson—not a member of the party—shot the fox

and claimed the corpse. Post sued for ownership. The court awarded ownership to Pierson, establishing (or reinforcing) the legal rule that (with respect to wild animals) ownership rights are bestowed on the person who first "possesses" the animal.

The "rule of first possession" is a common rule for resolving ownership disputes over previously unowned resources (Rose, 1985). The rule has been defended on efficiency grounds as a low-cost method of establishing rights, in that possession is often an unambiguous phenomenon (as in *Pierson v. Post*), providing a clear demarcation for establishing who owns a specific resource. However, the cost-benefit calculus does not always favor first possession, as issues of difficulty in determining possession, wasteful expenditures in the race to first possession, and possible disincentives to productive efforts, among other issues, may favor more finely tuned rules for establishing ownership to previously unowned resources (Lueck, 1995). Economic analysis of alternative rules for establishing initial ownership claims has focused on the cost of establishing such claims versus the benefits from the right to exclude (for an example with respect to the whaling industry, see Ellickson, 1989).

## The Prevention of Resource Depletion

It is a fundamental proposition in economics that an open-access resource will be used beyond the point of efficiency in the short run and quite likely to the point of depletion in the long run—the famous "tragedy of the commons" (Hardin, 1968). Hardin (1968) used the example of open grazing rights to village pastures that were a common arrangement in seventeenth-century England. The cost of using the pasture was zero to any individual herder; therefore, each herder had an incentive to graze his or her livestock as long as some return could be gained, that is, as long as some grass (or other grazing material) remained. This individual-maximizing behavior had two consequences. One, it reduced the current productivity of the pasture for other herders, as their livestock now had less grazing material. Two, it reduced the future productivity of the pasture, as it would be unable to regenerate itself.[2] This outcome results from rational individuals following their own self-interest and maximizing their individual productivity, a behavioral characteristic that in the world of Adam Smith results in beneficial social outcomes. Furthermore, it is of no consequence if a specific individual user (e.g., a herder) recognizes the seeds of the tragedy and attempts to avert the outcome by refraining from using the resource. His or her place would be taken by another user following his or her own maximizing instincts, and the tragedy would continue to its inexorable conclusion.

The crux of the tragedy lies not in the self-interested maximizing behavior of individuals but rather in the institutional setting in which such behavior is allowed to operate. In order to defeat the tragedy, an institutional setting must be found so that access to the resource can be limited. One possibility is to establish private ownership rights in the resources, with the accompanying ability to exclude. If the commons can be privatized, then maximizing incentives for the private owner will be in line with the socially efficient condition for resource utilization, as the profit-maximizing private owner will use the resource to the point where marginal revenue product equals marginal (factor) cost.

The tragedy of the commons, unfortunately, is no mere theoretical construct; rather, its destructive outcome is characteristic of some of the more serious problems currently facing the world's environment. Ocean fisheries serve as a textbook case of the tragedy of the commons in action, as open access has resulted in a state of near collapse for several species. The shrinking of tropical rain forests is another prominent example of the overexploitation of an open-access resource.

## Securing Returns From Investment

The act of investment requires incurring costs today in order to earn a return in the future. Such acts are unlikely without confidence in the ability to claim a future return, that is, without the ability to exclude others from the return to the investment. "Open access is associated with depletion and disinvestment rather than with accumulation and economic growth" (Eggertsson, 2003, p. 77). A system of private property rights can ensure that an owner who undertakes an investment today can secure the gains from that investment, thus eliminating the disincentive to investment that exists with an open-access resource.

The grant of exclusive rights to the use of the product of investment activities is to grant a form of monopoly power—the owner is the sole seller of the product. Such a grant creates no efficiency issues when the product is sold in a competitive market. Thus, if a wheat farmer is granted exclusive rights to the crop produced in his or her field, there will be no inefficiency if the wheat is sold in a competitive wheat market. The ownership right provides an incentive for efficient investment, and competitive markets will guarantee the efficient level of wheat production and consumption.

However, the scope of ownership rights—the power to exclude—can raise serious efficiency issues when such rights result in the creation of substantial monopoly power in product markets. These issues arise most notably when rights of exclusion are granted to the inventors of new products and the authors of creative works. This is the realm of intellectual property rights—patents and copyrights. The efficient level of output occurs when price, a measure of the social benefit from additional units of the good, equals the marginal cost of production, a measure of the opportunity cost to society of producing an additional unit of the good. A monopolist will set price higher than

marginal cost, thus denying access to the good for consumers whose benefit from additional units of the good (as indicated by the price they are willing to pay) exceeds the cost of producing additional units. Furthermore, monopoly prices will most likely result in extra-normal (monopoly) profits, resulting in a transfer of income from consumers to monopoly producers.

The economic justification for granting such monopoly power is that without the expectation of high returns, inventors and authors have little incentive to undertake the risky process of discovering and developing new products and expressive works. The benefits to society of having the new invention or creative work outweigh the costs of restricting its use. The specific provisions of patent and copyright law, such as the duration of the ownership rights and exceptions such as the "fair use" provision of the copyright law, attempt to strike a balance in the trade-off between incentives to create and monopoly inefficiencies. Changes in the parameters, which determine the magnitude of this trade-off, result in pressure to change the legal rules governing this area of property law.[3]

Recently, a more fundamental challenge has been leveled against the degree of monopoly power granted by patents and copyrights. Rather than promoting the creation of new products and creative works, the web of legal restrictions established by patent and copyright law may actually hinder discovery, innovation, and creative expression by creating substantial barriers to using the knowledge and ideas contained in existing protected works to develop new products and expressive works—an example of the "tragedy of the anticommons." The tragedy of the anticommons occurs when numerous individual property rights to a resource result in underutilization of that resource—the opposite outcome of the overutilization inherent in the tragedy of the commons (Heller, 1998).

## The Economics of Use Rights

If resources are scarce, then a society must answer the question of how those resources will be used. In a society characterized by private property rights, individuals who hold the ownership rights to the resources control the answer to that question, subject to the laws governing the use of their property. This freedom granted to private resource owners—"ownership sovereignty," if you will—reflects a fundamental tenet of economics—that individuals are the best judges of their own well-being and should be free to choose how to use their resources so as to maximize that well-being. Broad use rights conferred on individual owners are consistent with the underlying normative justification for a competitive market system—that the pursuit of individual self-interest in a competitive market system will result in a socially efficient allocation of resources.

### Externality Problem: When Use Rights Conflict

However, user rights are not unlimited. The conclusion that the decisions of rational, self-interested utility maximizers will result in a socially efficient allocation of resources rests on several assumptions. One key assumption is the absence of negative externalities. A negative externality exists when the activity of Person A imposes costs on Person B, and such costs are not reflected in the price Person A must pay to undertake the activity. It is a well-established principle in economics that the existence of negative externalities can result in a suboptimal allocation of resources.[4] Property law has likewise recognized the problems created by negative externalities, as perhaps best illustrated by the common law maxim "*sic utere tuo ut alienum non laedas,*" which translates to "use your own (property) in such a way that you do no harm to another's."

A straightforward example of the externality problem is presented by the case of *Orchard View Farm v. Martin Marietta Aluminum* (1980). The defendant's aluminum plant had emitted fluoride into the atmosphere, which damaged the plaintiff's orchards. From an economic standpoint, such emissions may be inefficient if the benefit from the emissions is less than the harm to the orchards (or, alternatively, the cost of reducing the emissions is less than the benefit to the orchard from reduced emissions). Assume that the harm to the orchards reduced the plaintiff's profits by $500, an amount that represents the net benefit to society (price paid by consumers minus cost of production) of the lost output. Assume further that the defendant could reduce the emissions to a nonharmful level by reducing output, with an accompanying loss of profits (representing again the net benefit to society of the foregone output of aluminum) equal to $300. Under this scenario, the reduction in aluminum output and the resulting increase in the output of the orchard would represent a more efficient allocation of resources, as the increased orchard output is more valuable to society than the reduction in aluminum output. However, the aluminum producer has no incentive to reduce output unless some corrective action is taken to "internalize" the external cost imposed on the orchard.

Property law provides one method to internalize the external costs—the award of compensatory damages for the harm done, often in a complaint of nuisance or trespass. (The *Orchard View Farms* case was a trespass action.) In the above hypothetical, if the court found the defendant liable for damages of $500, then the defendant would have an incentive to reduce output to the nonharmful level, as the $300 in lost profits would be less than the $500 in damages it would have to pay if it did not reduce output. The externality would be internalized, and the allocation of resources would be more efficient. (If the numbers in the above hypothetical were reversed, then the defendant would continue to produce at the same level and pay the damages, which would be the efficient result.)

The award of compensatory damages was one of several mechanisms that property law provided to deal with negative externalities. This mechanism was in line with traditional economic theory, which most often suggested the imposition of taxes (or fines) on the negative externality-generating activity. However, the publication of Ronald Coase's (1960) seminal article "The Problem of Social Cost" dramatically changed the way economists analyzed the question of negative externalities.

## The Coase Theorem

Instead of viewing negative externalities as a situation where Person A's activity imposed a cost on Person B, Coase (1960) recast the situation as one in which the two parties were vying over the use of a scarce resource in which no ownership rights had yet been established. If Person A were allowed to use the resource, then a cost would be imposed on Person B, who would not be able to use the resource. However, Coase argued that the converse was also true—if Person B were allowed to use the resource, a cost would be imposed on Person A, who would not be able to use the resource. In effect, externalities were reciprocal in nature, in that one party's use of the resource precluded the other party's use. In the *Orchard View Farms* case, the two parties were competing over the (incompatible) use of the atmosphere, to which neither (prior to the legal decision in the case) had an established property (use) right. The defendant's emissions imposed a cost on the plaintiff, but if the plaintiff were successful in preventing the defendant from releasing emissions, a cost would be imposed on the defendant. The question of which party is harming the other—that is, which party is creating the negative externality—cannot be determined until it is decided who has the right to use the resource in question.[5]

Drawing on this basic insight, Coase (1960) developed the famous Coase theorem: In the absence of transactions cost, the allocation of resources will be efficient regardless of the assignment of property rights. Placing the *Orchard View Farms* case in the idealized world of no transactions costs, the Coase theorem states that an efficient allocation of resources will result regardless of whether the plant is given the right to emit fluoride or whether the orchard is given the right to be free from fluoride emissions. To state the point more starkly, it does not matter, with respect to efficiency, what the court decides, as long as it decides something!

What is the logic behind this remarkable conclusion? It was concluded above that if the owners of the orchard were given the right to be free from harm from the emissions (via compensatory damages), the defendant would reduce output so as to avoid paying damages—the efficient outcome. If, instead, the owners of the plant were given the right to emit, then the orchard owners would have an incentive to offer the owners of the plant a payment, say $400, to reduce the emissions. The plant owners would agree to do so, as the $400 payment is greater than the $300 in decreased profit resulting from reducing emissions. As long as there are no obstacles to reaching this agreement—that is, there are no transactions costs—the efficient result is reached. The offer of payment by the orchard owners internalizes the cost of the emissions to the plant owners, as the refusal to accept the offer represents an opportunity cost to them.

This example illustrates the *strong* version of the Coase theorem, which states that allocation of resources will be *identical* (and efficient) regardless of the assignment of property rights. But identical allocational outcomes may not result, even in the absence of transactions costs. Efficiency requires that resources go to their most highly valued use. But "value" can be measured in two distinct ways—by how much one is *willing to pay* to acquire a resource (i.e., to buy) and by how much one is *willing to accept* to give up a resource (i.e., to sell). These two amounts may differ. It may require more to compensate a person for the loss of the right to a resource than the person would be willing to pay to acquire the right to that resource. For example, it may take $8,000 to compensate an individual who owns a beachfront cottage with beautiful sunset views to give up that right so that an oil well, which obstructs the view, can be drilled. However, that same individual, perhaps due to income constraints, may only be willing to pay $2,000 to prevent the oil well from being drilled. If there exists a divergence between willingness to pay and willingness to accept (i.e., if endowment effects exist), then the initial assignment (endowment) of the property right may result in a different allocation of resources.[6] However, regardless of the outcome, the resultant allocation of resources will be efficient, *given the property right assignment.* In other words, the resource will go to the most highly valued use regardless of the assignment of the right, but the most highly valued use may differ due to endowment effects stemming from the different assignment of the right.[7] This is the *weak* version of the Coase theorem: In the absence of transactions cost, the allocation of resources will be efficient (but not necessarily the same), regardless of the assignment of property rights.[8]

The Coase theorem is a fascinating and provocative proposition. Despite the fact that transactions costs are likely never zero, it has several implications for property law. One, the law should strive to reduce transactions costs so as to facilitate private agreements to resolve externality conflicts. A clear delineation of property rights is a key step in this process. Negotiations are likely to be particularly acrimonious if the parties believe their rights are being violated. Two, in cases where transactions costs are low, the courts do not have to burden themselves unduly worrying about the efficiency effects of their (clear) assignment of the property right in question, as bargaining will eventually result in an efficient outcome. The courts can focus on other possibly important aspects of their decision, such as the distributional or "fairness" implications. Third, when transactions costs are high, the court should

assign the resource in question to the most highly valued use. If transactions costs are high and the resource is assigned to the lower valued use, it is unlikely that a transfer to the more highly valued use will occur, thus resulting in an inefficient allocation of resources. This proposition is dependent on two key assumptions: one, that allocative efficiency is the primary objective of the legal system and, two, that the identification of the most highly valued use is clear and is not subject to possible endowment effects.

## The Economics of the Right to Transfer

The rules of property law are part of the set of rules that determine how resources can be passed from one user to another, that is, the right of alienability.[9] This transfer can include the transfer of all rights of use for all time or can be limited in both type of use and length of use. Full voluntary transference of rights can be done by sale, gift, or bequest. A partial transfer of rights can be done in a variety of legal forms, including licensing and leasing arrangements.

### Voluntary Transfers

It is a bedrock principle of economics that a voluntary transfer of resources between two reasonably informed parties improves the well-being of both parties, thus representing a Pareto superior reallocation of resources. This is perhaps the fundamental justification for a market-based economy. One prominent role of property law is to facilitate such transfers, primarily by clearly specifying property rights so that uncertainty over ownership can be removed from the market transaction.

But in some instances, the legal system explicitly restricts rather than facilitates the transfer of a property right. The restriction may be complete, as in the case of the right to vote, which can neither be sold nor given away. More often, however, the restraint on alienability is in the form of a prohibition on the sale of the resource, that is, the transfer of the resource in exchange for compensation. For example, it is illegal in most countries to sell one's bodily organs, such as a kidney, for purposes of transplantation, although the transfer of such via a donation is not only legal but often encouraged. Several economic justifications have been advanced in support of inalienability, among them the existence of negative externalities, poor information regarding the consequences of a sale, and myopic behavior (see Rose-Ackerman, 1985). While these justifications may be valid in certain situations, none address the fundamental question of why uncompensated transfers are allowed. The suspicion is that society fears that the offer of compensation will distort the owner's valuation of the resource—that is, that the lure of financial compensation will tempt the owner to transfer the resource when such transfer is not really in the owner's best interest. This is the classic paternalism argument. The argument is often applied most directly to lower income people, who may be particularly susceptible to an offer of financial compensation. Such temptation has been characterized in such adverse terms as "exploitation" of the poor, as if the provision of compensation is actually detrimental to the welfare of the poor. Paternalistic arguments violate the basic assumption in economics that individuals are the best judge of their own well-being, and thus restrictions on alienability based on such arguments are often criticized by economists as inefficient.

Questions have arisen as to exactly when and what ownership rights are given up in the process of a voluntary transaction, particularly when the transaction involves body tissue. In *Moore v. Regents of the University of California* (1990), the plaintiff agreed to the removal of his diseased spleen, which, unbeknownst to him, contained unique cells. Defendants, through genetic engineering, developed a highly lucrative line of pharmaceutical products based on the plaintiff's cells, without informing plaintiff. The patient claimed ownership; the defendant argued that all ownership rights to the spleen and cells therein were voluntary transferred at the time of the operation. The court ruled that the patient had forfeited his ownership right in the spleen once it was voluntarily removed from his body.[10]

### Involuntary Transfers

*Adverse Possession (Private Taking)*

Adverse possession is the rule of law that transfers ownership of property from the current owner to another private party, without the permission of or compensation to the owner. Adverse possession can be invoked when the owner has not objected to the "open and notorious use" of his or her property by another. For example, if a person builds a garage that encroaches five feet onto the adjacent property of another, and if, over an extended period of time, the owner of the adjacent property makes no attempt to claim ownership, then ownership of that part of the property may pass to the person who built the garage under a claim of adverse possession. The economic justification for such involuntary transfers is twofold. One, the property is moving to the person who values it more highly, on the presumption that the lack of objection by the owner indicates a very low valuation of the property by the owner. Two, adverse possession can discourage ownership claims based on occurrences in the distant past.

*Eminent Domain (Public Taking)*

Eminent domain is the power of the government to take private property to pursue a government function. The Fifth Amendment of the U.S. Constitution states, "No property shall be taken for public use without just compensation." Public works projects often require the acquisition of multiple parcels of private property, allowing the possibility of

one or more owners to "hold out" in an attempt to capture a larger share of the gains from the project. Assume that the government plans on building a hospital, with estimated net benefits (excluding property acquisition costs) of $100 million. The hospital is to be built on 10 identical parcels of private property, each with a market value of $1 million. Thus, the project is economically efficient, with net social benefits of $90 million. Assume the government engages in voluntary market transactions to acquire 9 of the 10 parcels at their aggregate market value of $9 million. The last parcel is now worth up to $91 million to the government, although no other buyer would pay more than $1 million. The owner of the last parcel is in a position to extract an enormous gain by refusing to sell—"holding out"—unless the government pays a substantial premium over the market price. The government may refuse to pay the premium, thus scuttling the project. If more than one of the owners of the 10 parcels attempts to follow this strategy, the problem is confounded. The power of eminent domain eliminates such strategic behavior, allowing the government to acquire property at (preproject) market value, which presumably represents the true opportunity cost of the property to society. The owner is left no worse off, and efficient projects can go forward.

Actual compensation is in accordance with the Pareto test for efficiency but is not required under the more widely used Kaldor-Hicks potential compensation test. However, the "just compensation" requirement promotes efficiency by forcing the government to more explicitly weigh the benefits of a proposed project against the actual costs, an incentive that is not as great if the government can simply take property without compensation. On equity grounds, the compensation requirement results in no one citizen bearing an unduly large cost of providing a public benefit—the general public will bear the cost through the taxation necessary to pay the compensation.

In principle, as well as practice, "fair market value" is considered the standard measure of "just compensation." However, market value—a willingness-to-pay measure of value—may not reflect the amount necessary to induce the owner to sell the property, that is, the amount necessary to fully compensate the owner for his or her loss. As discussed with respect to the Coase theorem, a property owner's willingness to accept may be greater than his or her willingness to pay. This is not unusual when the property in question is the owner's private residence—his or her home. In such a situation, the exercise of the power of eminent domain with "just compensation" based on fair market value may result in a reallocation that does not meet either the strict Pareto test for efficiency or even the weaker potential compensation test for efficiency.

The Fifth Amendment states that the government can acquire property for "public use." If title to the acquired property is transferred to the government and the property is available for use by the public, such as a park or a highway, the public use requirement is clearly fulfilled. The question has arisen, however, concerning the limits on the public use requirement. In *Kelo v. City of New London* (2005), the plaintiff challenged the city government's proposed taking of her house (with compensation at market value), the property to be converted to a parking lot for a private corporation in accordance with a comprehensive economic development plan formulated to reverse the city's declining economic fortunes. In a 5–4 decision, the Supreme Court ruled that the "public use" requirement was fulfilled, as the proposed development plan could provide substantial benefits to the general public.

### *Regulatory Takings*

The police power of the state enables the government to pass laws and regulations to protect the "health, safety, and welfare"[11] of the public. Such regulations may reduce the value of affected property, but the government is generally not required to compensate property owners for the reduced value. For example, the government can restrict the playing of outdoor music at a nightclub if the noise disturbs the peace and quiet of local residents without compensating the owner of the nightclub for any reduction in the value of his or her property resulting from reduced profits or even the closing of the business. Can such regulations so restrict the use of property that in effect the government has "taken" the property, even though title remains with the private owner? In *Lucas v. South Carolina Coastal Council* (1992), the defendant adopted a setback rule prohibiting construction on coastal property within a certain distance from the water, justifying the regulation as necessary to mitigate coastal erosion and to protect the coastal environment. The rule prevented the plaintiff from carrying out plans to build a house on each of his two coastline properties (purchased for $975,000). The plaintiff sought compensation for the significant reduction in the value of his properties, arguing that the regulation in effect constituted a "taking" under the Fifth Amendment. The Court held that if a new government regulation "deprived the owner of virtually all economic benefit" from his property and was not inherent "in the common law traditions of the state," then the regulation resulted in a government taking, requiring compensation under the Fifth Amendment.

The legal issue of whether a government regulation is a legitimate use of the state's police power with no compensation required for any resulting losses or whether the regulation is a taking requiring compensation remains unresolved. From an economic perspective, the analysis echoes the discussion above regarding the Coase theorem and the theory of externalities. As a logical proposition, it is equally correct to state, with reference to the *Lucas* case, that the property owner would harm the coastal environment if he built on the property (i.e., impose a negative externality) or, alternatively, that he would benefit the coastal environment if he did not build (i.e., confer a positive externality).[12] As with the Coase theorem, the assignment of property right must be decided (presumably the

role of the court) before the appropriate characterization of the externality can be determined.

## Protecting Property Rights

The primary remedies for the protection of property rights are injunctions and damages. An injunction is a court order requiring the defendant to cease the behavior that violates the plaintiff's property right in order to prevent "irreparable harm" to the plaintiff in the future. If the defendant violates the injunction, he or she can be held in contempt of court and face large fines or possible incarceration. The plaintiff can agree to have the injunction lifted if an agreement is reached with the offending party.

Damages are intended to compensate the holder of a property right for the damages that have resulted from the violation of his or her property right. As such, they are backward looking.[13] Damages and injunctions are not mutually exclusive remedies—a court may grant an injunction against future harms while awarding damages for past harms.

In their seminal article, Calabresi and Melamed (1972) highlighted a key difference between injunctions and damages as remedies. An injunction prevents another party from using (harming) the property of the owner (assuming that large fines and the threat of incarceration are effective deterrents). Damages, on the other hand, allow the offending party to use (harm) the property of another, as long as compensation is paid. Returning to the *Orchard View Farms* case, an injunction against emissions from the plant would result in the orchard being free from emissions. If damages were awarded (without an injunction), emissions could continue as long as the orchard owners were compensated for their losses. Calabresi and Melamed classified this distinction as protection by a *liability rule* (damages) and protection by a *property rule* (injunction).

Following the logic of the Coase theorem, in a world of zero transactions cost, the type of remedy should not affect the allocation of resources. Reversing the numbers in the *Orchard View Farm* hypothetical above, assume that profits from the plant are increased by $500 if emissions are released while damage to the orchard from the emissions equals $300. If the court awards damages, the plant owners will continue to allow the emissions and pay damages of $300, thus gaining $200. (Note that this represents an efficient allocation of resources.) On the other hand, if the court issues an injunction prohibiting future emissions, the plant owners have an incentive to offer the orchard owner an amount between $300 and $500 (say, $400) to lift the injunction and allow the emissions (again, the efficient outcome), an offer that the orchard owner would presumably accept. The choice of remedy does not affect the allocation of resources, only the distribution of the resultant efficiency gain.

As with the Coase theorem, the equivalency result relies on the key assumption that transactions costs are sufficiently low so that the two parties are able to reach an agreement. If transactions costs are high, then granting an injunction against the plant owners will result in the plant being forced to eliminate emissions, an inefficient outcome. In this situation, the award of damages (a liability rule) would lead to the efficient result, as the plant owners would continue to emit and pay the damages. As a general proposition, if transactions costs are low, an injunction is the preferred solution, as the parties can negotiate their own (efficient) agreement without the court undertaking the costly task of determining actual damages. However, if transactions costs are high, damages are the preferred solution so as to allow for the resource to be used by the more highly valued user, if that user is not awarded the property right.[14]

## Future Directions

The above discussion has touched on certain issues that will challenge the established rules of property law. Some form of property-like exclusionary rights must be established to currently open-access natural resources, or else such resources will soon face extinction. As new life-saving pharmaceutical products are developed, the ability of patent holders to restrict access via monopoly prices will become ever more costly to society. Medical advances in organ transplantation practices combined with increasing life expectancies of a healthier population will make the inalienability of bodily organs increasingly inefficient. The broader issue of property rights in one's bodily tissue, cell line, and genetic code will need to be addressed if the full benefits of scientific advances in biotechnology and genetic engineering are to be realized. In the area of government takings, the courts must decide how broadly the concept of "public use" will be defined and in what circumstances government regulations become so restrictive as to require compensation. But it should not be forgotten in the midst of these weighty national and global issues that one of the most important functions of property law is to decide conflicts over the appropriate uses of one's property that constantly arise at the most local of levels—disputes between neighbors.

## Conclusion

Economic efficiency requires that scarce resources are not exploited to the point of depletion, that they are employed in their most valuable use, and that productive investments are undertaken in order for the resource base to grow. Property law plays a key role in the successful obtainment of these objectives by establishing exclusion rights, use rights, and transfer rights to resources and then protecting those rights with appropriate remedies. Ownership rights grant to the owner of a resource the legal authority to exclude others from access (without permission) to the resource, thus providing an incentive against overutilization

and depletion as well as for productive investment in the growth of the resource. An owner has broad discretion to put the resource to the use that he or she deems most utility enhancing, thus promoting efficiency. However, that use may impinge on another's resource, and property law determines the extent to which an owner can use his or her resource in a manner that interferes with the property of another (the externality issue). By facilitating negotiations between the affected parties and, in situations in which negotiations are not feasible, by assigning use rights to the most highly valued use, the legal system can contribute to an efficient allocation of resources. A well-functioning system of property law will enable the voluntary transfer of resources between parties on mutually agreeable terms so that resources are moved to their most highly valued use. In situations where voluntary agreements may be difficult to achieve, the legal rules governing the involuntary transfer of ownership rights—adverse possession and government takings—can move resources to more highly valued uses, if properly structured. In a dynamic market economy, the effectiveness of established property law is constantly challenged by changing preferences of individuals, changing production technologies, and changes in the institutional arrangements governing society. The economic analysis of property law is aimed at evaluating how effective the current legal rules that govern ownership rights, as well as proposed changes to such, are in promoting an efficient and equitable allocation of resources in an ever-changing world.

## Notes

1. The Pareto test and the Kaldor-Hicks potential compensation test are the two standard concepts of economic efficiency (see Coleman, 1988).

2. More formally, an open-access resource will be used to the point at which the *average* revenue product of utilization equals the *marginal* cost of utilization, which is greater than the socially efficient level of utilization where *marginal* revenue product equals marginal cost.

3. For example, advances in file-sharing technology, which greatly lowered the cost of reproducing recorded music, have fueled the legal battle by the recording industry against the "pirating" (i.e., unauthorized reproduction) of copyrighted music (see *A&M Records v. Napster,* 2001).

4. Standard economic theory concludes that too many resources will be devoted to the negative externality-generating activity because individuals will not take account of the full cost of their activities, ignoring the external costs imposed on others.

5. If the reader is unconvinced by this reasoning, consider the following scenario. The aluminum plant has been a long-established business whose fluoride emissions created no harm on the surrounding people. The orchard owner purchases the land adjacent to the plant and proceeds to plant relatively delicate trees that are harmed by the emissions and then sues to have the emissions stopped. Whose actions impose a cost on whom?

6. The reader can trace out the different outcomes under an alternative assignment of the property right in the cottage owner/oil well example, assuming that the profits from drilling the oil well are $5,000.

7. This result is equivalent to the conclusion in welfare economics that a competitive market economy will generate an efficient allocation of resources, with the actual allocation dependent on the initial distribution of income and wealth (endowments). Change the initial distribution of endowments and the resulting allocation of resources will (presumably) change yet still be efficient

8. The distribution of economic welfare will, of course, change with the changing distribution of initial endowments.

9. Other fields of the law, most notably contract law and estate and trust law, play a prominent role in determining the conditions under which the transfer of ownership rights can occur.

10. The court did hold that defendant had violated other rights of the plaintiffs' (e.g., informed consent).

11. Protecting the "morals" of the public is sometimes added to this list of objectives for the police power.

12. In his majority opinion in the *Lucas* case, Justice Scalia stated, "A given restraint will be seen as mitigating 'harm' to the adjacent parcels or securing a 'benefit' for them, depending on the observer's evaluation of the relevant importance of the use that the restraint favors" (*Lucas v. South Carolina Coastal Council,* 1992, p. 1025).

13. "Permanent" damages, intended to compensate for anticipated future harms, can be awarded.

14. Kaplow and Shavell (1996) present a detailed economic analysis of the alternative remedies.

## References and Further Readings

*A&M Records v. Napster,* 239 F. 3d 1004 (2001).

Anderson, T. L., & McChesney, F. S. (Eds.). (2003). *Property rights: Cooperation, conflict, and law.* Princeton, NJ: Princeton University Press.

Barzel, Y. (1997). *Economic analysis of property rights* (2nd ed.). New York: Cambridge University Press.

Calabresi, G., & Melamed, A. D. (1972). Property rules, liability rules, and inalienability: One view of the cathedral. *Harvard Law Review, 85,* 1089–1128.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics, 3,* 1–44.

Coleman, J. (1988). *Markets, morals, and the law.* Cambridge, UK: Cambridge University Press.

Demsetz, H. (1967). Toward a theory of property rights. *American Economic Review, 57,* 13–27.

Eggertsson, T. (2003). Open access versus common property. In T. L. Anderson & F. S. McChesney (Eds.), *Property rights: Cooperation, conflict, and law* (pp. 73–89). Princeton, NJ: Princeton University Press.

Ellickson, R. C. (1989). A hypothesis of wealth-maximizing norms: Evidence from the whaling industry. *Journal of Legal Studies, 5,* 83–97.

Ellickson, R. C., Rose, C. M., & Ackerman, B. A. (2002). *Perspectives on property law* (3rd ed.). New York: Aspen.

Epstein, R. A. (1985). *Takings: Private property and the power of eminent domain.* Cambridge, MA: Harvard University Press.

Fischel, W. (1995). *Regulatory takings: Law, economics, and politics.* Cambridge, MA: Harvard University Press.

Gordon, H. S. (1954). The economic theory of a common property resource: The fishery. *Journal of Political Economy, 62,* 124–142.

Hardin, G. (1968). The tragedy of the commons. *Science, 162,* 1243–1248.

Heller, M. A. (1998). The tragedy of the anticommons: Property in the transition from Marx to markets. *Harvard Law Review, 111,* 621–688.

Hobbes, T. (1963). *Leviathan.* In J. Plamenatz (Ed.), *Thomas Hobbes: Leviathan.* Cleveland: The World Publishing Company. (Original work published 1651)

Kaplow, L., & Shavell, S. (1996). Property rules and liability rules: An economic analysis. *Harvard Law Review, 109,* 713–790.

*Kelo v. City of New London*, 545 U.S. 469 (2005).

Landes, W. M , & Posner, R. A. (2003). *The economic structure of intellectual property law.* Cambridge, MA: Harvard University Press.

*Lucas v. South Carolina Coastal Council*, 505 U.S. 1003 (1992).

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. (1995). An economic theory of adverse possession. *International Review of Law and Economics, 15,* 161–173.

Michelman, F. I. (1982). Ethics, economics and the law of property. In J. R. Pennock & J. W. Chapman (Eds.), *Ethics, economics and the law* (pp. 3–40). New York: New York University Press.

*Moore v. Regents of the University of California*, 793 P.2d 479 (1990).

Nash, C. A. (1983). The theory of social cost measurement. In R. Haveman & J. Margolis (Eds.), *Public expenditure and policy analysis* (3rd ed., pp. 56–79). Boston: Houghton Mifflin.

*Orchard View Farm v. Martin Marietta Aluminum*, 800 F. Supp (Oregon) (1980).

*Pierson v. Post*, 3 Cai. R. 175, 2 Am Dec. 264 (1805).

Rose, C. M. (1985). Possession as the origin of property. *University of Chicago Law Review, 53,* 73–88.

Rose-Ackerman, S. (1985). Inalienability and the theory of property rights. *Columbia Law Review, 85,* 931–969.

Sax, J. (1971). Takings, private property and public rights. *Yale Law Journal, 81,* 149–186.

Yandle, B. (1998). Coase, Pigou and environmental rights. In P. J. Hill & R. E. Meiners (Eds.), *Who owns the environment?* (pp. 119–152). Lanham, MD: Rowman and Littlefield.

# 75

# Queer Economics

## *Sexual Orientation and Economic Outcomes*

Richard R. Cornwall
*Middlebury College*

Queer economics looks at a particular example of the tight connection between markets and the emergence and shaping of social identities. What makes it "queer" is focusing on the role markets have played in the birth of lesbian-gay-bi-transexual identities with particular interest in how certain identities are queered (i.e., their social valuation is switched across a social boundary between straight-respectable and gay-abjected). Queer economics also goes the other direction—namely, from perceptions of sexual orientations in markets to the behavior of individuals and businesses in markets resulting, for example, in inequality in earnings according to sexual orientation.

## Conventional Economic Thinking

An economist ignorant of queer theory might imagine measuring the economic impact of queer culture on the circular flow of national output/consumption by measuring how many units of currency per year of queer culture are bought and sold. But then, does one count only sales of new lesbian, bisexual, gay, transgender, queer/questioning (LBGTQ) books, art, movies, television, radio, theater, receipts at gyms, Internet sex sites, sex clubs, and bars? Does one include phenomena deriving from LBGTQ cultures but aiming at mainstream culture such as dance clubs, Madonna videos, sexy couture, and even very mainstream clothing such as Abercrombie & Fitch's with a gay esthetic permeating its marketing through, especially, Bruce Weber's photos (McBride, 2005, chap. 2)? How should one account for designs created by LBGTQ people or sales by stores catering to LBGTQ consumers but also selling to straight folks?

This type of conventional thinking would see "the queer economy" as a distinct part of a nation's gross domestic product. Yet this approach fails since boundaries for LBGTQ cultures do not exist, and this conventional thinking is challenged by the new field of queer political economy (Cornwall, 1997). LBGTQ cultures are more like queer glasses: They transform how people view *all* culture since they change what is permitted, what is valued, what is disparaged, and how we conceptualize our economic lives.

We begin then by looking historically at the queering of social identities. This took place especially at the rollover from the nineteenth to the twentieth centuries simultaneously with the redefinition of boundaries between classes and genders. This *ver/mischung* of the codifications of class, sexuality, and gender is of fundamental importance for queer political economy.

## Part I: Markets Contribute to Creation of Identities Based on Sexual Orientations: The Rise of LBGTQ Identities

In the beginning, there were no sexual identities! Well, a bit more carefully, excluding earlier urban and cloistered, single-sex environments (e.g., see Boswell, 1980, on ganymedians in the eleventh century), there was no space in language and in thinking for what we label

LBGTQ. As D'Emilio (1983/1993) notes describing colonial North America,

> There was, quite simply, no "social space" in the colonial system of production that allowed men and women to be gay. Survival was structured around participation in a[n extended] nuclear family. There were certain homosexual acts—sodomy among men, "lewdness" among women—in which individuals engaged, but family was so pervasive that colonial society lacked even the category of homosexual or lesbian to describe a person. It is quite possible that some men and women experienced a stronger attraction to their own sex than to the opposite sex—in fact, some colonial court cases refer to men who persisted in their "unnatural" attractions—but one could not fashion out of that preference a way of life. Colonial Massachusetts even had laws prohibiting unmarried adults from living outside family units. (p. 470)

In the nineteenth century in the United States, a bit earlier in Britain, a bit later in some countries and continuing in many places today, has been a transformation from, on one hand, an economic system where most people live and work in kinship groups in agriculture to, on the other hand, the market system with individualized wage labor and with more urban living. This transformation was engendered by and simultaneously contributed to reduced costs of transportation and communication resulting in the spread of concentrated factory production sites and the spread of wage labor. This, in turn, enabled people to be able to conceive of being economically independent from their kin and from agricultural activities. It allowed LBGTQ individuals to move to urban areas and to meet each other and to discover their often hidden (from themselves) same-sex erotic interests as they participated as customers of boarding houses, molly houses, bathhouses, coffee shops, and cruising spots in, for example, England in the eighteenth century. Somewhat similarly, though changed by the filter imposed by gender in Western cultures, lesbians emerged as a self-aware and socially distinguished group in the nineteenth and twentieth centuries (see Cornwall, 1997, for references).

A tight cognitive link between biological sex, gender, and sexuality (Rubin's [1975] "sex/gender system"; see also Escoffier, 1985), from our vantage, appears to have been hegemonic until approximately the end of the nineteenth century. It was taken for granted, as "natural," that sexual object preference was determined by gender: "In the dominant turn-of-the-century cultural system governing the interpretation of homosexual behavior, especially in working-class milieus, one had a gender identity rather than a sexual identity or even a 'sexuality'; one's sexual behavior was thought to be necessarily determined by one's gender identity" (Chauncey, 1994, p. 48). In other words, the concept of gender consisted of the dichotomy of "male" versus "female," and this included sexual object choice. This tendency to focus on "gender" economized on cognitive categories, which simplified thinking by avoiding the ambiguities that can be expected in cases where there are small numbers of examples encountered by people as was certainly the case in rural settings.

The turn from the nineteenth to the twentieth century was a time not only of increasing urbanization in the United States but also of radical changes in the roles and extent of markets and the organization of production. The rise of wage labor in this country occurred early in the nineteenth century (e.g., "daughters of failing small farmers in the Northeast" began working in textile mills at Lowell; Amott & Matthaei, 1991, p. 295) and was followed in the second half of the century by dramatic growth of sex-segregated labor markets (Amott & Matthaei, 1991, pp. 315–348; Matthaei, 1982). Further and equally socially momentous was the rise of factories with thousands of workers under one roof, which led to enormous social upheaval as new codes for thinking about "productive" activities were developed.

## Interaction Between Shaping Class Boundaries and Sexuality Borders

This blender of changing social roles created a vortex of changing social identities:

> Working-class men and boys regularly challenged the authority of middle-class men by verbally questioning the manliness of middle-class supervisors or physically attacking middle-class boys. . . . [One contemporary] recalled, he had "often seen [middle-class cultivation] taken by those [men] of the lower classes as 'sissy.'" The increasingly militant labor movement, the growing power of immigrant voters in urban politics, and the relatively high birthrate of certain immigrant groups established a worrisome context for such personal affronts and in themselves constituted direct challenges to the authority of Anglo-American men as a self-conceived class, race and gender. (Chauncey, 1994, p. 112)

These struggles over where to map key social boundaries led

> politicians, businessmen, educators, and sportsmen alike [to protest] the dangers of "overcivilization" to American manhood. . . . Theodore Roosevelt was the most famous advocate of the "strenuous life" of muscularity, rough sports, prize-fighting, and hunting. . . . The glorification of the prize-fighter and the workingman bespoke the ambivalence of middle-class men about their own gender status . . . a "cult of muscularity" took root in turn-of-the-century middle-class culture. . . . Earlier in the nineteenth century, men had tended to constitute themselves as men by distinguishing themselves from boys. . . . But in the late nineteenth century, middle-class men began to define themselves more centrally on the basis of their difference from women . . . gender-based terms of derision [e.g., sissy, pussy-foot] became increasingly prominent in late-nineteenth-century American culture. (Chauncey, 1994, pp. 113–114)

This oversimplifies and ignores resistance to this respecification of gender ("They wonder to which sex I belong": Matthaei, 1995; Vicinus, 1992), but this recoding of masculinity seems to have been powerful at this time.

Closely tied to this redefinition of "male" in the 1890s was redefinition of class:

> Men and women of the urban middle class increasingly defined themselves as a class by the boundaries they established between the "private life" of the home and the rough-and-tumble of the city streets, between the quiet order of their neighborhoods and the noisy, overcrowded character of the working-class districts. The privacy and order of their sexual lives also became a way of defining their difference from the lower classes. (Chauncey, 1994, p. 35)

Just as a new "face" was being put on not-male (i.e., not-male became "female" instead of "boy"), so "middle-class" became "clean-face-and-well-laundered/mended-clothes" versus the "dirty" faces of slums. A quickly judged face was put on people living in slums: "The spatial segregation of openly displayed 'vice' in the slums had . . . ideological consequences: it kept the most obvious streetwalkers out of middle-class neighborhoods, and it reinforced the association of such immorality with the poor. . . . Going slumming in the resorts of the Bowery and the Tenderloin was a popular activity among middle-class men (and even among some women), in part as a way to witness working-class 'depravity' and to confirm their sense of superiority" (Chauncey, 1994, p. 26).

This simultaneous redefinition of gender, class, and occupations spilled over, "infected," the definition of sexual orientation that was occurring at the turn of the century:

> In a culture in which becoming a fairy meant assuming the status of a woman or even a prostitute, many men . . . simply refused to do so. . . . The efforts of such men marked the growing differentiation and isolation of sexuality from gender in middle-class American culture. . . . The effort to forge a new kind of homosexual identity was predominantly a middle-class phenomenon, and the emergence of "homosexuals" in middle-class culture was inextricably linked to the emergence of "heterosexuals" in the culture as well. If many workingmen thought they demonstrated their sexual virility by playing the "man's part" in sexual encounters with either women or men, normal middle-class men increasingly believed that their virility depended on their exclusive sexual interest in women. Even as queer men began to define their difference from other men on the basis of their homosexuality, "normal" men began to define their difference from queers on the basis of their renunciation of any sentiments or behavior that might be marked as homosexual. (Chauncey, 1994, p. 100)

Furthermore, "the queers' antagonism toward the fairies was in large part a *class* antagonism . . . the cultural stance of the queer embodied the general middle-class preference for privacy, self-restraint, and lack of self-disclosure" (Chauncey, 1994, p. 106).

Thus, it is no accident that this social earthquake of a lingual transformation in American culture, creating as "pervert" the distinct person now known as lesbigay, coincided with the rise of wage labor markets. John D'Emilio (1983) and Jeffrey Weeks (1979) have sketched this well for gaymen: how gender segregation in workplaces and in institutions for living and for social interaction such as clubs, baths, bars, and access to "public" spaces facilitated the evolution of notions of sexual identities. Especially important was the growth of wage labor in urban areas allowing gaymen to support themselves outside of traditional, kin-based agricultural networks,[1] and this, in turn, fed the rise of gay bars, baths, and so on, which, again in turn, made urban labor markets increasingly alluring for gaymen.

As Julie Matthaei (1995) notes, D'Emilio's "argument is much stronger for men than for women" (pp. 31–32, note 11; see also Chauncey, 1994, p. 27). The very different values that evolved for women compared to men in the last half of the nineteenth century in American culture (D'Emilio & Freedman, 1988) and the very different earnings levels resulting from the sex segregation of labor markets (Matthaei, 1995, p. 13) led to rather different manifestations of same-sex eroticism for women than for men. Thus, Matthaei (1995, pp. 12–14) offers tangible accounts of women whose transgendered performances in prominent public careers were "masterful" for their entire adult lives. Women passed as men, and their partnerships passed as marriage at a time when apparently few men exhibited similar reasons for living in drag. Analogously, while two women (neither transgendered) could be referred to as living in a "Boston marriage," no comparable term seems to have been used for two men living together. Thus, it seems important for socioeconomists to allow for the possibility of a (general equilibrium type of) simultaneity or interdependence in the social articulation of gender, sexuality, and labor and product markets.

What is "lesbian" and what is "gay" are fluid and are historically contingent on other social constructions. This poses a danger for economists who are as mentally conditioned as any other market players to seek discrete, firm economic identities that can be captured by yes/no decisions (zero/one dummy variables) across history. Although it may appear very likely (I conjecture, at our 2009 stage of imperfect "knowledge") that Sappho, Jane Addams, and Willa Cather may have shared a chromosomal structure differentiating them (with "statistical significance") from the chromosomal structures of more than 90% of the women who have lived on planet Earth, to then jump from conjectured or measured chromosomal patterns to inferences about the constructed trait now labeled "lesbianism," not to mention observed market behavior or even erotic activity, seems foolish if lesbianism and, indeed, sexuality in general are lingually based and are as fluid over time and place as lingual structures are easily observed to be. Thus, Jane Addams was able in the 1890s to exhibit market behavior

that a cliometrician might take for clear evidence of lesbian "identity"—that is, arranging in advance on her speech-making travels that each hotel provide a room with just one double bed for her and her "devoted companion," Mary Rozet Smith (Faderman, 1991, pp. 25–26)—yet a few decades later, Willa Cather kept her relation with her partner, Edith Lewis, of almost 40 years private until her death (O'Brien, 1987, p. 357). In between, the notion of "romantic friendship" had been replaced in Euro-based lingual cultures by the psychiatric diagnosis/identity-disorder of lesbianism.[2]

## Queer Theory's Perspective on Social Boundaries: Articulating What Is Unspeakable, Unthinkable

The term *queer theory* was first used by Teresa de Lauretis (1991) to describe "the conceptual and speculative work involved in discourse production, and . . . the necessary critical work of deconstructing our own discourses and their constructed silences" (p. iv). This focus on the use of language—on what is explicit and what remains hidden—studies people's discourse as a window into how these humans think and, especially, into how we (often unconsciously) categorize people and actions.

Humans depend on their linguistic communities to think, that is, to perceive, categorize, and articulate our desires (with erotic desires having an almost lexicographic priority). de Lauretis (1991) aimed to "problematize . . . to deconstruct the silences of history and of our own discursive constructions" (pp. iii, xvi). Judith Butler (1993) has noted,

> The construction of gender operates through *exclusionary* means, such that the human is not only produced over and against the inhuman, but through a set of foreclosures, radical erasures, that are strictly speaking, refused the possibility of cultural articulation. (p. 8)

The object of study in queer theory is the social articulation of same-sex eroticism and why, in recent centuries in Western-dominated cultures, this human interaction has been articulated as queer, as abject Other. The subtlety and complexity of this articulation led many, most notably Michel Foucault, who were searching for an analytical handle in this domain of socioeconomic inequality to the notion of discursive structure. I describe this as a mental structuring of concepts, each of which has an "aroma" of connotations, where these concepts are linked via physically developed neurological links that guide, often in a probabilistic and certainly in a nonconscious way (Damasio, 1994, p. 215), how we make inferences about what is "true"—hence, Foucault's (1972, p. 191; 1994, pp. 13, 68, 89) term *épistémè.* In short, discursive structures are (largely) linguistic cognitive structures—physically instantiated as ready-to-fire neural pathways in our brains—which develop as we learn our mother tongues and as we learn to understand, to map, our social embedment. Central to this queer thinking are the concepts of disgust, abjection, and Otherness.

## The Social Roles of Disgust, Abjection, and Otherness

> [I]n colonial America, [convicted] sodomites were more often than not lower-class servants, and the shoring up of patriarchal power was imbricated in nascent class divisions. One has only to look to other colonial situations of the time to see that that was not the only way the category of sodomy was being mobilized; the Spaniards, for instance, prone to see sodomites among the Moors in Spain, saw native cultures as hotbeds of irregular sexual practices. (Goldberg, 1994, p. 7)

Stallybrass and White (1986) note, "The bourgeois subject continuously defined and re-defined itself [as a way of distinguishing itself and its social legitimacy vis-à-vis the nobility and landed gentry] through the exclusion of what it marked out as 'low'—as dirty, repulsive, noisy, contaminating. Yet that very act of exclusion was constitutive of its identity. The low was internalized under the sign of negation and disgust" (p. 191). Slightly earlier, Stallybrass and White summarized this: "[Social] *differentiation . . . is dependent upon disgust*" (p. 191, emphasis added).

The ideas of abjection and of Otherness require careful explanation. The process of abjection gets started in the preverbal, deepest learning and mental formatting occurring when we are babies, and we develop perceptions of most-feared horrors that we may conflate with excrement, which is threateningly close to, even part of, ourselves. Since Freud and Lacan, there has been significant storytelling about such overarching psychic events shaping us. Julie Kristeva's (1982) formulation of how we transfer the symbolic and emotional meaning of these early experiences into verbal (and oneiric) articulations throughout the rest of our lives has been especially useful in queer theory.

Reading "Otherness" as mere difference, as simply being a mathematical reflection across an arbitrary and rather inconsequential boundary, is easy when transgenders such as RuPaul appear so sleekly, so *easily* on MTV. This is how most of the students in my queer studies classes who are not lesbigay first read "Other." Those who have not been subjected to the shame of being named faggot/dyke and of baring what they have (synecdochically) learned are their "most private," most individualizing parts of themselves before taunting gym classmates do not instantly jump to what queer scholars such as Judith Butler, Foucault, and de Lauretis or even writers such as Jean Genet and Dorothy Allison have grappled with. It is not enough to recite that the bloodiest hate crimes appear to be linked to the Levitical teaching about abomination, teaching that is carried out explicitly or implicitly by almost all churches

that, *at best,* merely tolerate those lesbigays who adopt the pose of synthetic straight (i.e., desiring "marriage" and all the other trimmings of liberal respectability we have inherited). It is not sufficient to point out how these teachings seem to encourage some people to follow a metanorm (Axelrod, 1986) to shoot/stab/bind-and-push-off-quarry-in-February-in-Vermont faggots and dykes. Reciting these episodes of brutality seems to communicate nothing of what Other describes.

So let's proceed didactically: "A *performative* is that discursive practice that enacts or produces that which it names" (Butler, 1993, p. 13). An example of a performative speech act is the creation of "currency": The inscription "This note is legal tender for all debts, public and private" is exactly what makes currency legal tender. Performatives gain collective credibility only through constant *reiteration.* Thus, currency *gains currency* only through its reiteration and its *anticipated* reiteration by juridical institutions and by private traders.

Another familiar performative is the utterance "I pronounce you man and wife." When certain institutionally designated people say this is a fact, then it becomes a fact, and it remains a fact to the extent that the husband and wife and their social interactions reiterate it. Similarly, "the norm of sex takes hold to the extent that it is 'cited' as such a norm, but it also derives its power through the citations that it compels" (Butler, 1993, p. 13). It does this by naming us, sexing us with culturally assigned connotations that place us in social space. Indeed, "the subject, the speaking 'I,' is *formed* by virtue of having gone through such a process of assuming a sex" (Butler, 1993, p. 3, emphasis added).

This assignment of gender roles gives us not only gender but also what we now call sexuality. This social process shaping female/male identities is the discursive means by which the heterosexual imperative enables certain sexed identifications *and forecloses and/or disavows other identifications.* This exclusionary matrix by which subjects are formed thus requires the simultaneous production of a domain of abject beings, those who are not yet "subjects" but who form the constitutive *outside* to the domain of the subject (Butler, 1993, p. 3).

"Abjection (in latin, *ab-jicere*) literally means to cast off, away, or out. . . . [T]he notion of *abjection* designates a degraded or cast out status within the terms of sociality. . . . [It] is precisely what may not reenter the field of the social without threatening psychosis, that is, the dissolution of the subject itself . . . ('I would rather die than do or be that!')" (Butler, 1993, p. 243, note 2).

In particular, Foucault (1990a and especially 1988 & 1990b) has sketched how Western discursive structures since late antiquity have very slowly evolved to make the male-female *couple* the social-civic atom (Foucault, 1988, p. 153). This evolution also made monogamy, "sexual monopoly" (Foucault, 1988, p. 149), a hegemonic doctrine and obliterated awareness—people stopped taking for granted—that there are good reasons for erotic intercourse other than procreation (Foucault, 1984/1990b, p. 181). Finally, Foucault (1978/1990a) most forcefully initiated linguistic study of the "exclusionary matrix by which subjects are formed" (i.e., the social process through which one's gender became more rigidly linked to the sex of her or his erotic partners).

Foucault (1978/1990a, p. 103) argued that this occurred through the development of the concept of sexuality. The breadth of Foucault's vision and of what might be involved in understanding the formation of social identities can be glimpsed from his careful analysis of what he saw as a change in the way "truth" is discovered/revealed (épistémè) from the seventeenth to the nineteenth centuries (Foucault, 1966/1994, 1972) *plus* his identification of "four great strategic unities which, beginning in the eighteenth century, formed specific mechanisms of knowledge and power centering on sex" (Foucault, 1978/1990a, p. 103) and were developed and "applied first, with the greatest intensity, in the economically privileged and politically dominant classes . . . the 'bourgeois' or 'aristocratic' family" (Foucault, 1978/1990a, p. 120).

1. "[T]he first figure to be invested by the deployment of sexuality, one of the first to be 'sexualized,' was the 'idle' woman" (Foucault, 1978/1990a, p. 121). This was the wife of the bourgeois market player.

2. The alarms about overpopulation economists associated with Robert Malthus brought about a "socialization of procreative behavior: . . . 'social' and fiscal measures brought to bear on the fertility of couples" (Foucault, 1978/1990a, pp. 104–105). This use of taxes/prohibitions to promote (for some governments) or restrain fecundity (some rapidly growing countries) continues to be a frequently employed and debated type of public policy.

3. "A psychiatrization of perverse pleasure: the sexual instinct was isolated as a separate biological and psychical instinct; a clinical analysis was made of all the forms of anomalies by which it could be afflicted" (Foucault, 1978/1990a, p. 105). Thus arose in the late nineteenth century the diagnosis of the pathological condition (identity) of being homosexual.

4. "As for the adolescent wasting his future substance in secret pleasures, the onanistic child who was of such concern to doctors and educators from the end of the eighteenth century to the end of the nineteenth, this was not the child of the people, the future worker who had to be taught the disciplines of the body, but rather the schoolboy, the child surrounded by domestic servants, tutors, and governesses, who was in danger of compromising not so much his physical strength as his intellectual capacity, his moral fiber, and the obligations to preserve a healthy line of descent for his family and his social class"

(Foucault, 1978/1990a, p. 121). Laqueur (2003, e.g., p. 13) has confirmed Foucault's view, noting how historically striking is this sudden social focus on masturbation: For well over a millennium and a half, masturbation was viewed as a minor theological infraction of good behavior, but in the eighteenth century, it suddenly reared up in Western social imagination as an especially heinous sin, perhaps the MOST heinous sin. It only gradually started to fade in importance in the twentieth century after the newly developed rainbow of erotic sins (the "psychiatrization of perverse pleasure"), ranging from homosexualities to diverse heterosexualities (heterosexuality was initially coined to label a particular variant of psychosexual "disease"), had grabbed the imagination of social facilitators: especially doctors and also, later, psychologists, reformers, ministers, and lawyers.

These newly positioned social facilitators, enunciators of *norms of respectability,* abjected so-called homosexual identities as exemplars of Other. The bourgeoisie "must be seen as being occupied, from the mid-eighteenth century on, with creating its own sexuality and forming a specific body based on it, a 'class' body with its health, hygiene, descent, and race" (Foucault, 1978/1990a, p. 124) This "queering" of certain behaviors and individuals, labeling them egregious transgressors of respectability, led those so labeled to react by adhering more strongly to each other as a distinct group, even flaunting the condemned behavior. Thus, the queer imprecations were queered by being embraced by their targets in newly evolving niches in labor, housing, and entertainment markets.

> Somewhat similar to the way in which, at the end of the eighteenth century, the bourgeoisie set its own body and its precious sexuality against the valorous blood of the nobles, at the end of the nineteenth century it sought to redefine the specific character of its sexuality relative to that of others, . . . tracing a dividing line that would set apart and protect its body. (Foucault, 1978/1990a, pp. 127–128)

Thus, Foucault conjectured that the social construction of the identities that we now term *lesbian, gay, bisexual,* and *transgendered* was part of, intimately tied to, the emergence of markets in Western societies. As noted above, George Chauncey (1994, pp. 13, 27, 111–126), Lillian Faderman (1991), John D'Emilio and Estelle Freedman (1988), and David Greenberg (1988) trace some of the details of this lingual construction of class boundaries as well as the coevolution of concepts of sexuality with institutions such as markets, churches, military training, and professional networks of doctors, social workers, artists, businesspeople, educators, professors, research people, and so on in the United States. This process leads to people generally investing heightened importance to categories we now label LBGTQ but had often hitherto seemed not worthy of much public notice.

## Part II: Perceptions of Sexual Orientations Affect Markets

This construction of LBGTQ identities is the lingual heritage of those growing up and learning to think in market-dominated cultures. This lingual inheritance determines in large part what we can think by determining the vocabulary of notions and connotations that are encoded in the neural circuits in our brains. It is these "flavors"/"smells" (i.e., connotations) of social identities that influence our perceptions of social labels and so enable us to articulate ideas about them. We grow axons and dendrites on neurons to connect neurons through their generation of and reception of neurotransmitters and thereby create these neural circuits (Panksepp, 1998, pp. 65, 85, 182–184). This physical embodiment of connotations in our brains makes the association of certain traits (e.g., "disgusting") with a social identity "automatic" and almost instantaneous. It precedes and provides the basis for "thinking," for deliberative cognition.

For just one example, this social language (and associated values and connotations) has made certain erotic desires and activities unspeakable, especially for politicians, and so has determined what data on erotic preferences and activity are (*not*) publicly funded and are (*not*) available as we seek to design public health measures to deal with human immunodeficiency virus (HIV) or to ascertain inequality in earnings (Badgett, 1995; Klawitter & Flatt, 1998).

## Inequality in Earnings by Sexual Orientation

Part I sketched the intensity of "antigay animus" (Badgett, 2007, p. 25), but is there any evidence that this animus actually affects markets? This evidence is important even if we accept the prevalence of antigay animus because it is often argued that markets can act to discipline would-be discriminators who would refuse to hire LBGTQ people or who would pay them less than other equally qualified people for the same work. Indeed, this is a standard economic argument since

1. the would-be discriminator is giving up profit by paying more than needed to accomplish the firm's hiring goals by not hiring less expensive and/or more qualified workers who happen to be LBGTQ and
2. competitive markets drive profits down to the minimum required to stay in business so this firm would be driven out of business.

This is the conventional neoclassical argument that, however, has been empirically refuted for, just for example, racial inequality, by Jim Heckman and Brook Payner's (1989) demonstration that passage of the 1964 Civil Rights

Act had an immediate effect on hiring black employees in textile mills in South Carolina. What blocks this neoclassical competitive-markets argument from dissolving inequality in employment is social norms that impose costs on those who violate them, as Heckman and Payner argued.

Badgett (2007, p. 27, Table 2.1) summarizes 12 econometric studies of inequality beginning with her pioneering work in 1995: "[Almost] every study using US data has found that gay/bisexual men earn less than heterosexual men, with a range of 13 percent to 32 percent" (p. 29). In particular, earnings for gay and/or bisexual men in the United States and the United Kingdom are estimated to range from 2% to 31% lower than for comparable straight men with similar human capital characteristics (education, age, region of country, partnership status, and, sometimes, occupation). Lesbians and bisexual women, on the other hand, in similar studies earn 3% to 27% more than straight women in most studies, although 3 of the studies found lesbians and bisexual women earning slightly less than straight women. As Badgett notes, "Lesbians do not earn less than heterosexual women, at least not when controlling for our imperfect measures of experience and human capital" (p. 32). And lesbians *do* earn less than straight men with similar productivity characteristics.

The preceding evidence of inequality tied to sexual orientation suggests that there might be an ameliorative role for antidiscrimination laws. The effects of local laws, however, are not entirely clear. The first study of them (Klawitter & Flatt, 1998) found "no impact on average earnings for people in same-sex couples" (Carpenter & Klawitter, 2007, p. 279) compared to straight couples. Carpenter and Klawitter's (2007) more recent study ends with a very weak conclusion on this potential ameliorative effect of legislation: "Policymakers should not abandon efforts to adopt and enforce policies that prohibit labor market discrimination against sexual minority individuals on the belief that they are ineffective" (p. 288). This tepidness results from their having found significant positive parameters for the effects of local antidiscrimination laws (in the presence, also, of a state law banning such discrimination) on gay-bisexual male earnings only for the effects of laws banning discrimination by *private employers* on the earnings of *government workers!* For female government workers, laws banning discrimination both by private employers and by public employers had a significant positive effect on earnings, but for both gay and lesbian employees of private employers, there was no significant effect.

What seems to be going on are two key selection effects:

1. Selection by LBG people to live/work in safer areas with protection against discrimination: A much higher fraction of the survey's LBG people (compared to straights) lived in localities with laws banning discrimination on the basis of sexual orientation (Carpenter & Klawitter, 2007, p. 283, Table 19.1), and also a higher fraction of LBG people lived in these areas than lived in areas with no laws banning discrimination. Badgett (2007) also notes that gay and bisexual men are choosing occupations where their coworkers will have less hostility toward them "or are going into more heavily female occupations than are heterosexual men" (p. 30).
2. Selection of which localities adopt laws banning discrimination against LBG people: Localities with nondiscrimination laws have higher earnings for all individuals, both straight and LBG, than do localities without such laws. This suggests that more prosperous and urban areas are more likely to adopt such laws and also provide a more welcoming locale for LBGTQ people (Carpenter & Klawitter, 2007, p. 285).

The work by Carpenter and Klawitter (2007) is one of the few studies using data from a large survey (California Health Interview Survey [CHIS] of 40,000 households in California in 2001 and 2003) where sexual orientation is determined by respondents' own reports ("Do you think of yourself as straight-heterosexual, gay (lesbian), or bisexual?" Carpenter & Klawitter, 2007, p. 281). This is one of the few direct sources of data on respondents' sexual orientation in the United States, so researchers have had to be very ingenious to tease this out of earlier data sources.

The first econometric work on LBGTQ inequality in earnings by Badgett (1995) used the General Social Survey (GSS) and inferred respondents' sexual orientation from their relative lifetime frequency of same-sex behavior since age 18. The GSS gives a rather small sample for each year, which is overcome by the 1990 and 2000 U.S. censuses where, however, there is no direct question on sexual orientation; rather, same-sex behavior is inferred for people who indicate they have an "unmarried partner" of the same sex. But even here, Badgett and Rogers (2002) found that 13% of couples in one survey of "cohabiting same-sex couples" (Badgett, 2007, p. 22) and 19% in another survey of such couples did not choose the "unmarried partner option" in the 2000 U.S. census, with this propensity not to respond being biased according to income: *Lower income couples were less likely to choose this option on the census.* This would tend to bias upwards resulting estimates of LBGTQ incomes, which means that the inequality in earnings according to sexual orientation is even *stronger* than indicated above, especially for gaymen.

The confusion on measuring the effects of local antidiscrimination laws might be due in part to the market impact of antigay animus being hidden by operating indirectly through a wage premium paid to married employees, all of which, until very recently, had to be in straight marriages. As Carpenter (2007) notes, "Employers who are uncertain about a worker's sexual orientation might plausibly use marriage as a signal for heterosexuality" (p. 77). He finds, again using CHIS data on cities in California, that the male marriage premium in California is large (18% after allowing for earnings differences due to age, ethnicity, education,

urbanicity, and occupation) and, indeed, it is largest in San Francisco, which has the highest percentage (28.4%) of adult men younger than age 65 say they are gay or bisexual. The premium is lowest in Riverside, which has the lowest percentage (4.5%) of adult men who are gay or bi. The marriage premium is also found to increase nonlinearly with age, as we would expect since an unmarried man age 50 is less likely to be straight than is an unmarried man age 20.

## Markets Segmented by Sexual Orientation

A response by profit-seeking business to the emergence of distinct LBGTQ people was to aim products (e.g., club clothes, bars and clubs, books, music and periodicals, Internet sites) to them. This formation of queer market niches gave these businesses a degree of brand-name distinctness and hence market power by which to earn monopoly profits. Of course, this evolution of markets reinforces the social cohesion of self-identifying participants in queered markets, but it can also seduce non-queer, self-imagined non-homophobic people ("metrosexuals") as well as not-gay-self-identified LBGTQ people who discover a certain pleasure (perhaps of tweaking norms of respectability) gotten when consuming these newly queered goods (e.g., going to LBGTQ dance clubs). This evolution suggests to marketers-producers that they can enlarge profits by aiming at self-conceived "sophisticated" heterosexuals.

Thus, we might argue that LBGTQ identities get doubly queered by markets: They act as a social blender by mixing and matching body-costume-identity parts across whatever social identities currently exist. This both reifies boundaries between identities (e.g., queering previously straight products like music/dance venues) and also dissolves these cultural boundaries ("queering" LBGTQ product niches by seducing non-LBGTQ people to join LBGTQ consumers and so making this product less queer). Thus, queering is like negation: Double queering results in un-queering.

## Conclusion

Queer political economy shares with feminist and race theory interest in the social articulation of cognitive codes (what in psychology are termed *schemas*) that stigmatize bodies with certain traits and so amplify social inequality. This interest in the perception of bodies differs from both neoclassical analysis and classical Marxian analysis, which have constructed analytical methods that ignore "desiring bodies" and instead model the interaction in markets of bodiless actors whose "desires" have been largely erased. See Cornwall (1997) for detail on the erasure of preferences from economics.

Microeconomic analysis that ignores the interactions between social labeling and the operation of markets distorts our economic policy making, including, for example, marketing, land use planning, and fecundity projections. In particular, it blinds us to substantial inequality in earnings and occupational choice and the evolution of market niches by sexual orientation.

## Notes

1. See also Chauncey (1994), who notes that urban areas also offered "relatively cheap accommodations and the availability of commercial domestic services for which men traditionally would have depended on the unpaid household labor of women" (p. 135).

2. Cather indicated sensitivity to this lingual transformation occurring during her young adulthood: "When [Cather] confessed to Louise [Pound, her first big love while a student at the University of Nebraska] that she thought it unfair that feminine friendships were 'unnatural,' she used a loaded word which reveals that certain intense female friendships were being defined as deviant—as lesbian—by the dominant culture. In the 1890s, *unnatural* was a code word . . . that meant 'deviant,' 'aberrant,' or 'homosexual'" (O'Brien, 1994, pp. 58–59).

## References and Further Readings


Amott, T., & Matthaei, J. (1991). *Race, gender, and work: A multicultural economic history of women in the United States.* Boston: South End.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review, 80,* 1095–1111.

Badgett, M. V. L. (1995). Wage effects of sexual orientation discrimination. *Industrial and Labor Relations Review, 48,* 726–739.

Badgett, M. V. L. (2007). Discrimination based on sexual orientation: A review of the literature in economics and beyond. In M. V. L. Badgett & J. Frank (Eds.), *Sexual orientation discrimination: An international perspective* (pp. 19–43). New York: Routledge.

Badgett, M. V. L., & Frank, J. (Eds.). (2007). *Sexual orientation discrimination: An international perspective.* New York: Routledge.

Badgett, M. V. L., & Rogers, M. (2002). *Left out of the count: Missing same-sex couples in census 2000.* Amherst, MA: Institute for Gay and Lesbian Strategic Studies.

Blandford, J. (2003). The nexus of sexual orientation and gender in the determination of earnings. *Industrial and Labor Relations Review, 56,* 622–642.

Boswell, J. (1980). *Christianity, social tolerance, and homosexuality: Gay people in Western Europe from the beginning of the Christian era to the fourteenth century.* Chicago: University of Chicago Press.

Butler, J. (1993). *Bodies that matter: On the discursive limits of "sex."* New York: Routledge.

Carpenter, C. (2007). Do straight men "come out" at work too? The heterosexual male marriage premium and discrimination against gay men. In M. V. L. Badgett & J. Frank (Eds.), *Sexual orientation discrimination: An international perspective* (pp. 76–92). New York: Routledge.

Carpenter, C., & Klawitter, M. (2007). Sexual orientation-based antidiscrimination ordinances and the earnings of sexual

minority individuals. In M. V. L. Badgett & J. Frank (Eds.), *Sexual orientation discrimination: An international perspective* (pp. 277–292). New York: Routledge.

Chauncey, G. (1994). *Gay New York: Gender, urban culture, and the making of the gay male world, 1890–1940.* New York: Basic Books (HarperCollins).

Cornwall, R. R. (1997). Deconstructing silence: The queer political economy of the social articulation of desire. *Review of Radical Political Economics, 29,* 1–130.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain.* New York: Putnam.

de Lauretis, T. (1991). Queer theory: Lesbian and gay sexualities. An introduction. *differences: A Journal of Feminist Cultural Studies, 3,* iii–xviii.

D'Emilio, J. (1983). Capitalism and gay identity. In A. Snitow, C. Stansell, & S. Thompson (Eds.), *Powers of desire: The politics of sexuality* (pp. 100–113). New York: Monthly Review Press. (Reprinted in *The lesbian and gay studies reader,* by H. Abelove, M. A. Barale, & D. Halperin, Eds., 1993, pp. 467–476. New York: Routledge)

D'Emilio, J., & Freedman, E. B. (1988). *Intimate matters: A history of sexuality in America.* New York: Harper & Row.

Escoffier, J. (1985). Sexual revolution and the politics of gay identity. *Socialist Review, 15,* 119–153.

Faderman, L. (1991). *Odd girls and twilight lovers: A history of lesbian life in twentieth-century America.* New York: Columbia University Press.

Foucault, M. (1972). *The archaeology of knowledge and the discourse on language* (A. M. Sheridan Smith, Trans.). New York: Pantheon.

Foucault, M. (1988). *The care of the self: Vol. 3. The history of sexuality* (R. Hurley, Trans.). New York: Vintage (Random House). (Original work published 1984)

Foucault, M. (1990a). *The history of sexuality: Vol. I. An Introduction.* New York: Vintage (Random House). (Original work published 1978)

Foucault, M. (1990b). *The use of pleasure: Vol. 2. The history of sexuality* (R. Hurley, Trans.). New York: Vintage (Random House). (Original work published 1984)

Foucault, M. (1994). *Les mots et les choses: Une archéologie des sciences humaines* [The order of things: An archaelogy of the human sciences]. New York: Vintage (Random House). (Original work published 1966)

Frank, J. (2007). Is the male marriage premium evidence of discrimination against gay men? In M. V. L. Badgett & J. Frank (Eds.), *Sexual orientation discrimination: An international perspective* (pp. 93–103).New York: Routledge.

Goldberg, J. (Ed.). (1994). *Reclaiming Sodom.* New York: Routledge.

Greenberg, D. F. (1988). *The construction of homosexuality.* Chicago: University of Chicago Press.

Heckman, J. J., & Payner, B. S. (1989). Determining the impact of federal antidiscrimination policy on the economic status of blacks: A study of South Carolina. *American Economic Review, 79,* 138–177.

Klawitter, M. M., & Flatt, V. (1998). The effects of state and local antidiscrimination policies on earnings for gays and lesbians. *Journal of Policy Analysis and Management, 17,* 658–686.

Kristeva, J. (1982). *Powers of horror: An essay on abjection* (L. S. Roudiez, Trans.). New York: Columbia University Press.

Laqueur, T. W. (2003). *Solitary sex: A cultural history of masturbation.* New York: Zone Books.

Matthaei, J. (1982). *An economic history of women in America: Women's work, the sexual division of labor, and the development of capitalism.* New York: Schocken.

Matthaei, J. (1995). The sexual division of labor, sexuality, and lesbian/gay liberation: Towards a Marxist-feminist analysis of sexuality in U.S. capitalism. *Review of Radical Political Economics, 27,* 1–37.

McBride, D. (2005). *Why I hate Abercrombie & Fitch: Essays on race and sexuality.* New York: New York University Press.

O'Brien, S. (1987). *Willa Cather: The emerging voice.* New York: Oxford University Press.

O'Brien, S. (1994). *Willa Cather.* New York: Chelsea House.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions.* New York: Oxford University Press.

Rubin, G. (1975). The traffic in women: Notes on the "political economy" of sex. In R. R. Reiter (Ed.), *Toward an anthropology of women* (pp. 157–210). New York: Monthly Review Press.

Stallybrass, P., & White, A. (1986). *The politics and poetics of transgression.* London: Methuen.

Vicinus, M. (1992). "They wonder to which sex I belong": The historical roots of the modern lesbian identity. *Feminist Studies, 18,* 467–498.

Weeks, J. (1979). *Coming out: Homosexual politics in Britain, from the nineteenth century to the present.* New York: Quartet Books.

# 76

# Economics and Religion

Carmel U. Chiswick
*University of Illinois at Chicago*

The last two decades of the twentieth century saw an explosion of empirical as well as theoretical research into the relationship between religion and economic behavior. For the most part, this research ignores theological differences, focusing instead on behavioral differences associated with different religious identities. The causation runs both ways: Some studies analyze the effects of religious identity on various economic activities, and others analyze the effects of economic incentives on religious observances and institutions. Both of these lines of research have yielded strong results and have dramatically affected our understanding of the relationship between economics and religion. Prices and incomes are powerful incentives that invariably influence the actions of individuals, and the human capacity for creative rationalization contributes to the widespread evasion of costly behaviors, including costly religious strictures.

Before economics became a modern social science, casual observation generated many stereotypes about differences between religious groups regarding economic success, differences that were often attributed to differences in religious teachings. Today these arguments are viewed with skepticism. Some are based on stereotypes that do not stand up to empirical scrutiny. Others are based on an imperfect understanding of the religious teachings to which they refer. Recent research suggests that some of the most important differences between religious groups can be explained not directly, by the religious strictures themselves, but indirectly by intervening variables that affect the economic incentives faced by individuals.

To provide an overview of this subject, this chapter begins with a consideration of the economic incentives affecting a consumer's decisions in the religious marketplace—that is to say, the demand for "religion." It will then look at how this demand affects religious institutions and generates a supply of religious goods and services. Other topics will include the structure of this religious marketplace, the related "marketplace for ideas" in a religiously pluralistic society, and religious human capital. Finally, there will be a brief discussion of empirical findings for the effects of religious affiliation and intensity of belief or practice on selected economic behaviors.

## The Demand for Religion

From the perspective of an individual consumer, religious expression is an economic good that must compete with all other goods for a share of the resource budget. It is not a good in the material sense but rather an intangible for which people express a preference by their willingness to spend time and money on its acquisition. Nor is it a good that can be purchased in a consumption-ready form. It belongs to the category of economic goods that must be self-produced by each individual. The consumer may buy goods and services that contribute to this end but must spend his or her own time to use them in a way that creates a religious experience.

The theological aspects of any particular religion may be thought of as its technology, a set of "recipes" or blueprints for behaviors, expenditures, and beliefs that will produce the desired results. This gives the consumer a production function that converts time and money—that is, labor and capital—into an output that can be called "religious experience." The substance of this theology is generally irrelevant for an economic analysis, much as the theory of the firm can analyze its behavior without specifying the particular good it produces or the specifics of its

production. What follows here is a similar abstraction, an analytical framework that can be applied to good effect for deepening our understanding of many different religious behaviors.

In their seminal work on this subject, Azzi and Ehrenberg (1975) suggest that religion is best thought of as a bundle of three distinct but interrelated goods. First there is spirituality, the desire for which seems to be a primal human impulse that finds some sort of expression in every society from the earliest times to the present. Then there is the fact that religion always seems to have a collective dimension—an individual "joins" or "belongs to" a particular religion and observes various rituals as a "member" of this group, typically in conjunction with other adherents. These two aspects of religion are sometimes referred to as its "spiritual good" and its "social good," respectively. Religion also addresses the dilemma of human mortality, the frightening inevitability of death and its implications for the meaning of life. This is usually referred to as the "afterlife good," although not every religion speaks to this need by positing an explicit life after death. In most religions, these three "goods" are bundled into a single product called "religion," but since their economic attributes differ, it is useful to consider them separately.

## Supernatural Being(s): The Spiritual Good

For many people, the search for spirituality is at the heart of any religious experience. Not everyone feels deeply about this—preferences vary, just as they do when the subject is ice cream or toys or fashionable clothing. Regardless of the priority placed on it, however, for most people spirituality is the central quality that effectively defines an experience as religious.

Spirituality is a classic example of a self-produced good. It is very, very personal and can never be acquired without intimate involvement of the individual consumer. The religious technology (theology) provides a guide for behaviors that will achieve this, religious rituals and their associated objects are designed to facilitate the process, and religious professionals are there to support and direct these activities. Yet in the end, it is the individual alone who has this spiritual experience at a deeply personal level. Like other dimensions of human capital, it cannot occur without the individual's participation, and once it has occurred, it is inalienable from that person.

Although the spiritual impulse seems to be a basic human need, the extent to which a person chooses to indulge in it is certainly affected by its price. Using a basic two-factor production function approach, the full price of this good is the direct cost of purchased goods and services and the indirect cost of the time spent in pursuit of spirituality. Some religions make heavy demands on consumers' incomes, but many of the most popular can be practiced with little direct expenditure. The search for spirituality thus tends to be a relatively time-intensive activity, and its full price is therefore sensitive to the value of a consumer's time.

Time is valued at its opportunity cost, and a consumer's budget is allocated optimally when the marginal value of time is the same in every alternative use. It is conventional to use the wage rate as a first approximation of this value, whether the actual wage for those who participate in the labor force or the shadow wage for those who do not. The full price of a time-intensive religious activity is thus positively related to the wage rate. Full income is also positively related to the wage rate, especially for people in the positively sloped region of their labor supply curve. The effect of higher wages on the demand for spirituality is thus ambiguous: A higher full price reduces the quantity demanded, but the higher income shifts the demand curve to the right. Empirical studies suggest that for most Americans, the price effect dominates the income effect so that higher wages are associated with less time spent in religious activities seeking spirituality.

In the production process, a relatively high cost of labor is an incentive to "economize" by becoming more capital intensive. In the search for spirituality, this takes the form of religious practices that substitute money for time. A high-wage person, for example, might purchase expensive religious objects but spend little time using them, might donate generously to causes associated with godliness, or might hire a substitute to engage in specific religious rituals on his or her behalf. Expensive time is also an incentive for innovations that raise its marginal product, whether by investing in skills relevant for the production of spirituality or by altering the religious environment in ways more suitable for (complementary to) the reduced time inputs.

This model has implications that result in testable hypotheses that appear to be consistent with the behavior of American consumers. Wealthy consumers often donate large amounts of money to religious causes even though they may not devote much of their own time to religious activities. Congregations with less time-intensive religious practices, like shorter services or fewer holy days, tend to attract disproportionately congregants at the upper end of the wage distribution. American religious institutions have also been innovative in adapting to the spiritual needs of consumers with a high value of time, for example, with services conducted in the English language or sermons applicable to a busy lifestyle.

## Belonging: The Social Good

In contrast with the search for spirituality, which is an intensely personal activity, adopting a specific religion implies participating in a group of similarly inclined individuals. This aspect of the demand for religion is analyzed as a "club" good, drawing on an extensive economics literature on club theory. Like other self-produced goods, a club good cannot be purchased directly but must be produced

with the consumer's own time and effort. A club good, however, cannot be produced by a single consumer in isolation. The productivity of resources that one individual devotes to making this good depends on the resource allocations made by other members of the club. For example, joining the church choir has different implications for a consumer's religious experience depending on how many others join the choir and with what intensity of religious participation.

Although all religions contain some measure of this characteristic, they vary in the way in which it is displayed. At one extreme, it may be possible to "buy" a membership, either directly or indirectly by making a large donation. Such a group would lack spiritual content and thus raises the question as to whether it is truly a religion. Most religions, however, require some participation in group rituals related to worship, to life cycle celebrations, or to obtaining or demonstrating merit by performing good deeds. In each case, the religious experience a consumer obtains as output depends not only on the effective use of his or her own resources but also on their complementarity with the resources devoted by other participants in the group.

Because of this interaction, clubs are a "quasi-public" good in the sense that they have some but not all properties of a public good. Like a true public good, a consumer can belong to a religion without diminishing its availability to other consumers. Unlike a true public good, however, a club can devise means of limiting membership and thus excluding potential consumers. This can be done by charging a membership fee or by specific criteria such as age, gender, race, profession, national origin, or place of residence. Although some religious groups use such means of restricting entry, these are generally eschewed by most American religions on the basis of theological, social, or political principles.

A club with important interpersonal complementarities that does not limit entry is faced with the classic "free-rider" problem. Since the productivity of a consumer's resources is enhanced by the resources devoted to the club by its other members, he or she has an incentive to choose a group where the other members spend more than he or she does. In effect, individuals try to economize on their own resources by substituting the resources of others. But it is mathematically impossible for everyone in the group to spend below the average. People spending more than the average are getting less output for their resources and have an incentive to seek another group where they could obtain a greater benefit from the same resource expenditure. When these people leave the original group, it begins an immiserating spiral that makes it increasingly unable to attract new members.

In a classic paper, Iannaccone (1992) considers this free-rider problem in the context of religious groups. He points out that many religions impose implicit taxes on their members as a means of supporting the religious group itself. This can take the form of tithing, of requiring the purchase of expensive religious articles, or of social pressure to donate money. There can also be a "tax" on time if membership requires volunteering for time-intensive ritual or charitable activities. Even in the absence of such taxes, however, many religions require a "sacrifice" of goods or time by which is meant a donation that is actually destroyed as part of a religious ritual. A sacrifice does not contribute directly to the support of the group itself, but from the individual's perspective, it is similar to a membership fee. It thus serves to discourage people from joining if their resource contributions would otherwise be lower than the value of the required sacrifice. By discouraging participation by people whose commitment to the group is low, a large sacrifice can raise the average level of commitment and thus benefit the remaining group members indirectly. A religion may also impose nonmonetary requirements to discourage adherents who might otherwise leave the group. Requirements such as those affecting clothing, appearance, or diet serve to identify adherents as committed members of the group but would be stigmatizing in the world of nonadherents. Both sacrifice and stigma are commonly observed characteristics of religion that serve to limit the problem of free riders in the religious community.

## An Unusual Investment: The Afterlife Good

Mortality is at the heart of the human condition, and religion is an important way in which people deal with the uncertainties and loss associated with their own death and that of their loved ones. Religions typically address this issue by embedding the relatively short life span of a human being in a larger picture of eternal life. There are various ways in which a theology deals with this question, but one that is very common is to posit a more or less explicit life that a person will experience after his or her own death. As long as an action during a person's current life on earth will have consequences for his or her circumstances in this afterlife, there is an incentive to alter current behavior with a view toward this long-run future. The benefits of good behavior induced by this theology are summarized by the term "afterlife" good.

A religious theology posits afterlife rewards to people who spend their time and money on "good" behaviors and afterlife punishments to those who spend their resources on "bad" behaviors. To the extent that this causes a consumer to alter his or her spending patterns, it trades present utility for future rewards after death. In this respect, it is best thought of as an investment rather than a consumption good, and as such, it can be treated analytically like any other investment. Other investments, however, are typically designed to yield their rewards at a later point in a person's lifetime, whereas the afterlife good pays off only after the investor is dead. The optimal strategy is thus to invest first in prospects that mature earlier and postpone this late-payoff investment until later in life. This is consistent with the observation that older people tend to spend more time and

money on religious participation than do youngsters, and it reinforces any tendency for people to become more concerned with the afterlife as they face their own mortality.

## The Supply of Religion

Since religious experience is a self-produced good, there is no explicit market for it and so no supply curve in the usual sense. Yet there is a market for religious goods and services, and there is a large sociology literature that views the religious sphere as having a "marketplace of ideas" (Warner, 1993). In this marketplace, religious groups compete with each other for adherents in much the same way that firms compete for customers, and individuals seek out a religious congregation to join in much the same way as they shop for other goods and services. Much of this sociology literature is concerned with the structure of this market, highly competitive in the United States but more like a monopoly in countries with a state religion. This analogy has contributed much to a new understanding of the economic aspects of religion.

A fundamental requirement for a market to be competitive is free entry of new firms and free exit of firms that are unsuccessful. Religious "startups" are characteristic of the American religious scene. New congregations frequently appear within established religious denominations, and entirely new religions can and do emerge. Most of these new religions are small, and many of them eventually disappear for lack of followers, but some—like the Church of the Latter-day Saints (Mormon) and Christian Science—have been very successful and grew into established religions.

Unlike religious monopolies that are licensed (and funded) by the government and typically managed by a religious hierarchy or bureaucracy, competitive religious markets are characterized by independent congregations that hire their own clergy. The clergy in a competitive market are responsible directly to their congregants and thus tend to be more sensitive to their religious needs. Also unlike a religious monopoly structure, congregationalism finds it more efficient to conduct nonritual religious functions (e.g., charities or proselytizing) in a separate set of para-religious organizations. It is also common for congregations within the same religious denomination to form an umbrella organization (analogous to an industry group) to represent their common interests in the larger society.

Each of these types of religious organization is characteristic of the United States, a pluralistic society in which religious markets are highly competitive. In countries with one or more state religions, however, the government-sanctioned religious body typically carries out all of these religion-related functions. As an indication of how distinctive American religious pluralism actually is, the separation of function associated with religious pluralism is frequently described as a symptom of "Americanization" in a religious group.

## Religious Human Capital

Most people think of themselves as having been born into a religion, suggesting that perhaps they have no choice as to where they belong. While it is true that a person may be born into a family that practices a certain religion, it is not true that this religion is innate in a newborn child. In fact, religious education and training are an important part of a child's upbringing, often from a very young age. The consequence of this training is that youngsters accumulate human capital—skills, knowledge, memories, sensations—specific to a particular religion, denomination, or perhaps even congregation. The more religious human capital a person has, the more efficiently he or she can obtain a religious experience from any given amount of resources. Religious education is an important activity for the community as a whole as well as for its individual members, and it is the core function of any proselytizing undertaken by a religious group.

A human capital approach provides additional insights into the workings of a competitive religious market for adults. Each religion may be thought of as having its own religion-specific human capital, the formation of which is characterized by the usual positively sloped marginal cost curve. Each person may be thought of as accumulating religion-specific human capital until the (shadow) marginal rate of return to religious education approaches the marginal rate of return to other types of human capital. If a person were to convert to a different religion, the human capital specific to the "old" religion would lose its productive value, and human capital specific to the "new" religion would need to be acquired. The economics of religious switching is formally analogous to occupational change or to international migration, a new investment that would be attractive only if the benefits outweigh the costs. The incentives are such that conversion is less likely to occur the greater the human capital intensity of either the "old" or the "new" religion, and it is most likely to occur between denominations with similar human capital where religious skills are highly transferable as, for example, among mainline Protestant denominations in the United States. Religious switching is also more likely among young adults who have not yet made heavy religion-specific investments.

### Religion and Socioeconomic Behavior

Religions differ with respect to the compatibility of their teachings with other aspects of the society to which their adherents belong. This can be analyzed as the degree of complementarity between religious and other forms of human capital and the mutual complementarity among

different kinds of human capital investments (Chiswick, 2006). People whose religious teachings complement the public school curriculum, for example, would have higher rates of return to both types of education and therefore an incentive to invest in both religious and nonreligious human capital. Adherents of these religions tend to have high levels of education, better health, lower fertility, and marriage patterns that tend to go along with these attributes.

In contrast, people whose religious teachings are anti-complementary (i.e., contradictory) to a public school curriculum would have an incentive to specialize in either religious or nonreligious investments in human capital. Those who invest more heavily in religious human capital would face lower rates of return to investments in secular education, for example, and those who choose to invest in nonreligious forms of human capital would have less incentive to invest in religious education. In these denominations, adherents who are very religious tend to have low levels of education and health, high fertility, and marriage patterns associated with their consequent low socioeconomic status, and adherents with greater secular achievements would tend to have lower levels of religious observance.

## Empirical Evidence

Empirical analyses of economic and demographic behaviors in the United States suggest that religion is an important factor in many decisions related to education, health, fertility, marriage, and divorce. This literature distinguishes between religious affiliation, on one hand, and the degree of religiosity, on the other. Whether or not a person identifies himself or herself as belonging to a particular religion or denomination seems to be less important than the intensity of religious observance and the degree of commitment to the group. Some of these findings fit conventional stereotypes, but some do not.

### Data on Religion and Economics

Empirical analyses of the effect of religion on economic and demographic behaviors are constrained by the paucity of data. Data collected by the U.S. government generally do not have a question on religion. A few economic surveys ask respondents to self-identify as Protestant, Catholic, Jew, or Others, categories that are too heterogeneous for testing hypotheses about religious behavior. The National Survey of Religious Identity (NSRI) and the American Religious Identity Survey (ARIS) have nationwide random samples with considerable detail for self-identified religion. A number of other sets of data are available from the Web site of the Association of Religion Data Archives (ARDA; http://www.thearda.com), which also have questions about religion, some more detailed than others, but few of these data sources have information on employment or wage rates that would be useful to test economic hypotheses. Newer surveys are beginning to address this problem, but in the meantime, only a few of the existing data sets can be used to study the influence of economic factors on religious behavior.

Studies that use economic data to analyze the effects of religious identification on economic and demographic behaviors find that the usual religion categories—Protestant, Catholic, Jew, Other—are too heterogeneous to be very useful. The aggregation principles for religious groupings should be analogous to those used for aggregating factors of production or industrial output. Religions can be grouped together into a single category if they are close substitutes with each other or if their respective types of religious human capital have similar complementarity properties with respect to nonreligious human capital. Religions should be separated into different groupings if neither of these conditions holds. The so-called Mainstream Protestant denominations can be grouped together because they typically have very similar religious human capital. Fundamentalist Protestant denominations can be grouped together because they typically share a strained relationship between religious human capital and some of the nonreligious human capital of mainstream America. In contrast, Mainstream Protestants and Fundamentalist Protestants should not be grouped with each other because they differ with respect to both religious human capital and its complementarity with nonreligious human capital.

Empirical results are much clearer when religious identification variables are classified according to these principles. It has become conventional to distinguish between "fundamentalist" and "mainstream" Protestant denominations. If the data permit, it is also useful to split the Mormons and the "African" Protestant denominations into separate categories. The Other category includes a number of very small groups, but whenever possible, the people reporting no religion (including agnostics and atheists) should be separated from those identifying with small religious groups (e.g., Greek Orthodox, Buddhists, Hindus, Moslems).

Variables relating to the degree of religiosity—that is, to the intensity of religious observance without regard to the particular religion—also need to be interpreted with caution. Some of the most common questions ask about the frequency of attendance at religious services and donations of money (and sometimes of time) to religious organizations. Other questions may ask about beliefs: whether there is a supernatural deity (God), whether there is life after death, or whether the words of the Bible are to be taken literally. These questions are reasonably good indicators of religiosity for Protestant religions and perhaps for Christians in general, who usually comprise the large majority of American respondents. For other religions, however, they may be less apt. The concept of God, for example, may be different for some of the Asian religions

than it is for the monotheistic religions of Judaism, Christianity, and Islam. As another example, intensely religious Jews may interpret the Bible literally only in its original Hebrew language, subject not only to variations in translation but also to a variety of possible interpretations of its original intent. Such differences reduce the effectiveness of these questions as general indicators of religiosity, although the problems are assumed to be small for samples with mainly Christian respondents.

As an increasing number of immigrants bring with them a religion that is relatively new to the United States, another issue arises with regard to intensity of religious practice. This occurs when a religion is specific to a particular ethnic group, as is the case for Jews, Greeks, Armenians, and Russians. (It was also the case for Roman Catholics in an earlier era, when immigrants from Ireland, Italy, Germany, and Poland attended separate churches, but their descendants today are no longer deeply divided along ethnic lines.) The distinction between ethnicity and religion is not always clear in such cases, and survey respondents might indicate belonging to a religion when in fact their identity is primarily with the ethnic group. Even if their actual beliefs are similar to those of agnostics or atheists, the fact that they observe religious rituals as part of their ethnic activity may lead them to self-identify otherwise.

### Some Preliminary Findings

With these considerations in mind, a number of empirical studies have investigated the effects of religion and religiosity on economic and demographic behaviors (Lehrer, 2009). The evidence for the United States is generally consistent with the predictions of economic theory, but for the most part, the particular religion to which a person belongs does not seem to matter as much as the fact that a person belongs to some religion rather than none. It is possible that this arises because people ignore religious teachings (theology) when making human capital investment decisions. It is possible, however, that in a pluralistic society, religion would have low explanatory power for statistical reasons. For example, suppose people tend to choose an affiliation compatible with their nonreligious human capital portfolio, and suppose religious groups tend to adapt practices and even teachings to be compatible with the characteristics of their members. The data would then show that people are sorted into religious groups by their socioeconomic characteristics, and there would be little additional explanatory power for religion after controlling for the usual determinants of a human capital investment.

The empirical evidence for the United States also suggests that the degree of religiosity has a very important effect on investments in nonreligious human capital. Measures of religiosity describe an individual's commitment to religious practices (e.g., church attendance) and the intensity of his or her belief in its theology. For at least some of the socioeconomic outcomes, religiosity seems to have a nonlinear effect. For example, education levels tend to rise with religiosity up to a point, but people with very high levels of education tend to have low levels of religiosity. This pattern is consistent with predictions of the human capital model outlined above. People whose religious human capital is complementary with secular investments would exhibit a positive relationship between religiosity and, say, education, while those with anti-complementary religious human capital would combine high religiosity with low education levels or low religiosity with high education levels.

## Institutional Change

Americans affiliate with religions that have adherents in other parts of the world. Some of these are international, with a leadership established somewhere in its "world headquarters," while others are rooted in a single country to which adherents look for inspiration and guidance. In either case, however, the adherents living in the United States typically alter their observances (and even sometimes their beliefs) to better fit the American socioeconomic scene in a process that is labeled "Americanization." This may be perceived as a falling off of religious observance, yet the evidence suggests that Americans are among the most religious people in the modern world.

Economic analysis suggests an alternative interpretation in which Americanization is seen not as rampant materialism but rather an adaptive response to different economic circumstances. High American wage rates provide an incentive to substitute goods for time in the production of any religious experience—hence the observed tendency toward "materialism"—but they also provide an incentive to improve the efficiency of time spent in religious observance. In a competitive religious marketplace, people seek the religious community most compatible with their personal preferences, and clergy have an incentive to be sensitive to the religious needs of their congregants. Religious education also adapts to the relatively high education level of American congregants, and human capital formation is another means of raising the efficiency of time spent in religious activity. As Americans adapt their consumption patterns to changes in their economic environment, their religious consumption patterns and even theologies also change, and their congregations change along with them.

This institutional adaptability goes a long way toward explaining why religion continues to play an important role in American life despite all predictions to the contrary. Karl Marx characterized religion as "the opiate of the masses" that should disappear with economic development—it has not. Others believed that religion could not survive the scrutiny of science and would disappear among people with high levels of secular education—it has not. Instead, evidence for the United States suggests that when wages and wealth are held constant, religious participation actually

increases with the level of education. By placing religion and religiosity in their economic context, it is possible to obtain a deeper appreciation of the social importance of religion and its ability to thrive in many different circumstances.

## References and Further Readings


Azzi, C., & Ehrenberg, R. (1975). Household allocation of time and church attendance. *Journal of Political Economy, 83,* 27–56.

Chiswick, C. U. (2006). An economic perspective on religious education: Complements and substitutes in a human capital portfolio. *Research in Labor Economics, 24,* 449–467.

Iannaccone, L. R. (1990). Religious practice: A human capital approach. *Journal for the Scientific Study of Religion, 29,* 297–314.

Iannaccone, L. R. (1992). Sacrifice and stigma: Reducing free-riding in cults, communes, and other collectives. *Journal of Political Economy, 100,* 271–291.

Kosmin, B. A., & Keysar, A. (2006). *Religion in a free market: Religious and non-religious Americans.* Ithaca, NY: Paramount.

Lehrer, E. L. (2009). *Religion, economics, and demography: The effects of religion on education, work, and the family.* New York: Routledge.

Warner, R. S. (1993). Work in progress toward a new paradigm for the sociological study of religion in the United States. *American Journal of Sociology, 98,* 1044–1093.

# 77

# Economics and Corporate Social Responsibility

Markus Kitzmueller
*University of Michigan*

Corporate social responsibility (CSR) constitutes an economic phenomenon of significant importance. Today, firms largely determine welfare through producing goods and services for consumers, interest for investors, income for employees, and social and environmental externalities or public goods affecting broader subsets of society. Stakeholders often take account of ethical, social, and environmental firm performance, thereby changing the nature of strategic interaction between profit-maximizing firms, on one hand, and utility-maximizing individuals, on the other hand.

Hence, CSR is referred to as "one of the social pressures firms have absorbed" (John Ruggy, qtd. in *The Economist,* January 17, 2008, special report on CSR) and considered to "have become a mainstream activity of firms" (*The Economist,* January 17, 2008; Economist Intelligence Unit, 2005). Many (inter)national firms strive to achieve voluntary social and environmental standards (e.g., IS014001), and the number of related certifications in Organisation for Economic Co-operation and Development (OECD) countries as well as in emerging market economies is constantly growing. Broad access to the Internet as well as comprehensive media coverage allow the public to monitor corporate involvement with social ills, environmental degradation, or financial contagion independent of geographical distance. A 2005 U.S. survey by Fleishman-Hillard and the National Consumers League (*Rethinking Corporate Social Responsibility*) concludes that technology is changing the landscape in which consumers gather and communicate information about CSR and that Internet access has created a "more informed, more empowered consumer . . . searching for an unfiltered view of news and information."

In light of (a) such "empowered" market participants able to discipline firms according to their preferences and (b) the public good nature of business "by-products," policy makers must reevaluate the border between public and corporate social responsibility. In this context, Scherer and Palazzo (2008) note that "paradoxically, today, business firms are not just considered the bad guys, causing environmental disasters, financial scandals, and social ills. They are at the same time considered the solution of global regulation and public goods problems" (p. 414). In sum, CSR opens up a wide array of economic questions and puzzles regarding firm incentives behind voluntary and costly provision of public goods as well as the potential welfare trade-off between their market and government provision. While economic research had initially addressed the question of whether CSR possesses any economic justification at all, it has recently shifted to how it affects the economy, stressing the need of analytical machinery to better understand the mechanisms underlying CSR as well as its interaction with classical public policy. Therefore, the objective of this chapter is to identify, structure, and discuss essential economic aspects of CSR.

At first sight, CSR appears to be at odds with the neoclassical assumption of profit maximization underlying strategic firm behavior. Corporate social performance often means provision of public goods or reduction of negative externalities (social or environmental) related to business conduct. As public goods and externalities entail market failure in the form of free riding or collective action problems,

government provision through direct production or regulation may be most efficient, a concept generally known as Friedman's classical dichotomy. If firms still decide to engage in costly social behavior beyond regulatory levels (i.e., CSR), then why would they voluntarily incur these costs, and is this behavior overall economically efficient?

The attempt to answer these questions leads to the firm's objective—maximizing shareholder value—and its dependence on the nature of shareholders' and stakeholders' preferences. Shareholders and investors can be profit oriented and/or have social and environmental preferences. The same is true for consumers, while workers may be extrinsically and/or intrinsically motivated. This heterogeneity in preferences (i.e., the presence of nonpecuniary preferences alongside classical monetary ones) is able to shed light on CSR within standard economic theory.

Another important issue intrinsically related to CSR concerns information asymmetries between firms and stakeholders. Reputation and information are important determinants of consumer, investor, employee, or activist behavior and, therefore, firm profits. Hence, many firms proactively report on their CSR activities and consult governments, international organizations, nongovernmental organizations (NGOs), and private auditors to earn credibility. In short, firms seek to build and maintain social or environmental reputation in markets characterized by information asymmetry and socially or environmentally conscious agents.

While information economics, contract, and organization theory provide a suitable framework to analyze the motivations and strategies beneath CSR, public economics and industrial organization may enhance the understanding of how the underlying "social pressures" might affect market structure, competition, and total welfare. The remainder of this chapter is organized as follows: The second section defines CSR and discusses the classical dichotomy between the public and private sectors in light of CSR. The third section outlines the crucial role of preferences in explaining and conceptualizing CSR. The fourth section gives a structured overview of distinctive theoretic explanations of strategic CSR in light of some empirical evidence. The fifth section concludes.

## What Is CSR? From Definition to Analysis

Before entering economic analysis, the stage has to be set by defining *corporate social responsibility*. In practice, a variety of definitions of CSR exists. The European Commission (2009) defines corporate social responsibility as "a concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders on a voluntary basis." The World Bank (n.d.) states,

> CSR is the commitment of businesses to behave ethically and to contribute to sustainable economic development by working with all relevant stakeholders to improve their lives in ways that are good for business, the sustainable development agenda, and society at large.

A notion similar to "voluntary behavior" can be found in definitions that refer to either "beyond compliance," such as those used by Vogel (2005) or McWilliams and Siegel (2001), who characterize CSR as "the fulfillment of responsibilities beyond those dictated by markets or laws," or to "self-regulation," as suggested by Calveras, Ganuza, and Llobet (2007), among others.

These attempts to define CSR reveal two basic conceptual features: First, CSR manifests itself in some observable and measurable behavior or output. The literature frequently refers to this dimension as corporate social or environmental performance (CSP or CEP). Second, the social or environmental performance or output of firms exceeds obligatory, legally enforced thresholds. In essence, CSR is corporate social or environmental behavior beyond levels required by law or regulation. This definition is independent of any conjecture about the motivations underlying CSR and constitutes a strong fundament for economic theory to investigate incentives and mechanisms beneath CSR. Note that, while Baron (2001) takes the normative view that "both motivation and performance are required for actions to receive 'the CSR label,'" it is proposed here that linking a particular motivation to the respective performance is required for the action to receive "the correct CSR label" (e.g., strategic or altruistic). From an economic point of view, the "interesting and most relevant" form of CSR is strategic (i.e., CSR as a result of classical market forces), while McWilliams and Siegel's (2001) definition would reduce CSR only to altruistic behavior.

The logical next step is to build the bridge from definition to economic analysis. CSR often realizes as a public good or the reduction of a public bad. Hence, revisiting the classical dichotomy between state and market appears to be important. Relevant works that relate CSR with public good provision include Bagnoli and Watts (2003) and Besley and Ghatak (2007), who explicitly define CSR as the corporate provision of public goods or curtailment of public bads. Firms may produce a public good or an externality jointly with private goods, either in connection with the production process of private goods (e.g., less polluting technology such as in Kotchen [2006], or safe/healthy working conditions) or linked to the private good/service itself (e.g., less polluting cars or energy-saving light bulbs). This perspective on CSR relates directly to earlier work by J. M. Buchanan (1999), who referred to such joint provision of a public and private good as an "impure public good," and relevant insights such as those derived by Bergstrom, Blume, and Varian (1986) in their seminal paper on the private provision of public goods, which can be readily translated into the CSR framework. For example, Bergstrom et al. focused on the interaction between public and private (individual) provision of the public

good and the effect on overall levels of provision and concluded that public provision crowds out its private counterpart almost perfectly. Along these lines, Kotchen (2006) compares joint corporate provision of private and public goods in "green markets" (where the private good is produced with an environmentally friendly production technology) and separate provision of either, leading to the similar conclusion that the very same crowding out takes place between corporate provision and individual provision and may even lead to an overall reduction in the level of the public good. More precisely, Besley and Ghatak (2001) notice that public goods provision has dramatically shifted from public to mixed or complete private ownership in recent years, while Rose-Ackerman (1996) phrases the problem as the "blurring of the analytically motivated division between for-profit, nonprofit and public sectors in reality." To explain these observations, Besley and Ghatak suggest that in the presence of incomplete contracts, optimal ownership is not a question of public versus private provision but simply should involve the party that values the created benefits most. Another interesting rationale provided by Besley and Ghatak (2007) identifies economies of scope (i.e., natural complementarities between private and public goods production, leading to cost asymmetries/advantages on the firm side) to be the decisive variable in determining the efficiency of impure public goods. The conclusion states that if economies of scope are absent, tasks should be segregated into specialized organizations (i.e., governments provide public goods and firms private ones). Otherwise, CSR might very well be optimal if governments or not-for-profit providers are unable to match CSR levels of public good provision due to opportunism, cost disadvantages (= economies of scope argument), or distributional preferences. All these findings are of immediate importance to those authorities involved in the mechanism design of public good provision.

Assuming that private and public good production is naturally bundled, the major trade-off between government regulation and CSR can be summarized as follows: While government regulation of firms may entail the production of optimal or excessive levels of public goods, the allocation of costs and benefits may be suboptimal due to the uniformity of public policy tools (i.e., firms have to charge higher prices and cannot sell to those consumers without sufficient willingness to pay for the impure public good anymore; this also denies those consumers the acquisition of the pure private good under consideration). On the other hand, CSR may achieve second best levels of the public good combined with distributional optimality inherent in the working of markets. Under special circumstances (e.g., if a government foregoes regulation because an absolute majority of voters does not have preferences for the public good), CSR can even Pareto improve total welfare by serving the minority of "caring" consumers (Besley & Ghatak, 2007). In sum, policy makers should take into account the systemic constraints of both public and corporate provision of public goods.

While analyzing CSR through a "public economics lens" offers important insights into welfare implications, efficiency, and comparative and normative questions regarding CSR and public policy, it does not shed sufficient light on the motivations behind CSR. Therefore, the next section develops a categorization of CSR along motivational lines and across theoretical frameworks. In short, CSR can be subclassified as either *strategic,* market-driven *CSR,* which is perfectly compatible with profit maximization and Milton Friedman's view of the *socially responsible firm,* or as not-for-profit CSR that comes at a net monetary cost for shareholders. However, foregone profits (note that Reinhardt, Stavins, & Vietor [2008] define CSR in this spirit as *sacrificing profits in the social interest*) due to costly CSR need not be at odds with the principle of shareholder value maximization and do not automatically constitute moral hazard by managers if shareholders have respective intrinsic (social or environmental) preferences that substitute for utility derived from extrinsic (monetary) sources. Hence, any microeconomic explanation of CSR builds upon the recent advancement of new concepts of individual behavior in economics and the related departure from the classical *homo oeconomicus* assumption. In other words, the economic rationalization of CSR is closely linked to the extension of traditional individual rational choice theory toward a broader set of attitudes, preferences, and calculations.

## From *Whether* to *Why:* Economics and the Evolutionary Understanding of CSR

Initial research into CSR was dominated by the question of *whether* firms do have any social responsibility other than employing people, producing goods or services, and maximizing profits. However, firms increasingly engaged in CSR activities that, at first sight, seemed to be outside its original, neoclassical boundaries. Hence, research shifted focus to why firms actually do CSR. Both questions, whether and why CSR, are intimately related and will be jointly addressed in this section.

Should firms engage in CSR? And if so, why (not)? In this respect, Milton Friedman (1970) examined the doctrine of the social responsibility of business and concluded that the only responsibility of business is to maximize profits (i.e., shareholder value), while goods or curtailment of bads based on public preferences or social objectives should be provided by governments endowed with democratic legitimation and the power to correct market inefficiencies (such as free riding or collective action problems). Based on the assumption of perfect government, this view suggested that CSR was a manifestation of moral hazard by managers (firm decision makers) toward shareholders and not only inefficient but also inconsistent with the neoclassical firm's profit orientation. But rather than putting

the discussion about CSR to a halt, Friedman's thoughts provoked the search for an economic justification of CSR in line with neoclassical economics. The breakthrough came with the idea that CSR may actually be a necessary part of strategy for a profit-maximizing firm. In other words, profit maximization can be a motivation for CSR.

But how may CSR be integrated into the objective function of the profit-maximizing firm? The answer to this question builds upon the existence of preferences that are beyond those of the classical *homo oeconomicus.* Stakeholders as well as shareholders often are socially or, in general, intrinsically motivated, a fact that profit-maximizing firms cannot ignore as it directly affects demand in product markets and/or supply in labor markets. Furthermore, such preferences may induce governments to intervene in the market via regulation or taxation while fishing for votes (as stakeholders are at the same time voters and thereby determining who stays in power/government). In sum, social stakeholder preferences translate into some sort of action or behavior relevant to corporate profits. Therefore, CSR qualifies as part of a profit-maximizing strategy. CSR induced by demand side pressures or as a hedge against the risk of future regulation has been termed *strategic CSR* by Baron (2001), while McWilliams and Siegel (2001) refer to the same underlying profit orientation of CSR as a *theory of the firm perspective.*

If shareholders have preferences allowing them to derive intrinsic utility equivalent to extrinsic, monetary utility, any resulting social or environmental corporate performance will constitute a nonstrategic form of CSR that is equally consistent with Friedman's (1970) view of the firm. Here, the objective of the firm reflects the preferences of its owner(s) and therefore might involve a reduction of profits or even net losses without breaking the rule of shareholder value maximization. So Friedman's concept of CSR being equal to profit maximization has been confirmed and enriched by taking account of a new set of stakeholder and shareholder preferences. The result is a bipolar conception of CSR being either strategic or not for profit with varying implications for the financial performance of a firm. Figure 77.1 summarizes the four basic combinations of stakeholder and shareholder preferences and their implications for CSR. If shareholders are purely profit oriented, the firm should act strategically, maximize profits, and engage in CSR efforts only if stakeholders demand it. On the other hand, if shareholders care about corporate environmental and social conduct, CSR will always act as a corporate channel of contributing to public goods independent of stakeholder preferences. In this case, profit maximization is not the target, and nonstrategic firms may forego profits or incur losses to be borne by shareholders. The size of these losses depends, however, on stakeholders' willingness to pay for and general attitude toward CSR.

At this point, a general discussion of the crucial role of individual preferences in the economic analysis of CSR is in order.

It was again Friedman (1970) who explicitly pointed out that to understand any form of social responsibility, it is essential to notice that *society is a collection of individuals and of the various groups they voluntarily form* (i.e., incentives, preferences, and motivations of individual share- and stakeholders determine organizational behavior). Stiglitz (1993, 2002) talks about new concepts to be taken into account when modeling individual behavior. Becker (1993) proposes an "Economic Way of Looking at Behavior," stressing the importance of a richer class of attitudes, preferences, and calculations for individual choice theory. What Friedman, Stiglitz, and Becker have in mind is a new class of psychological and sociological ideas that

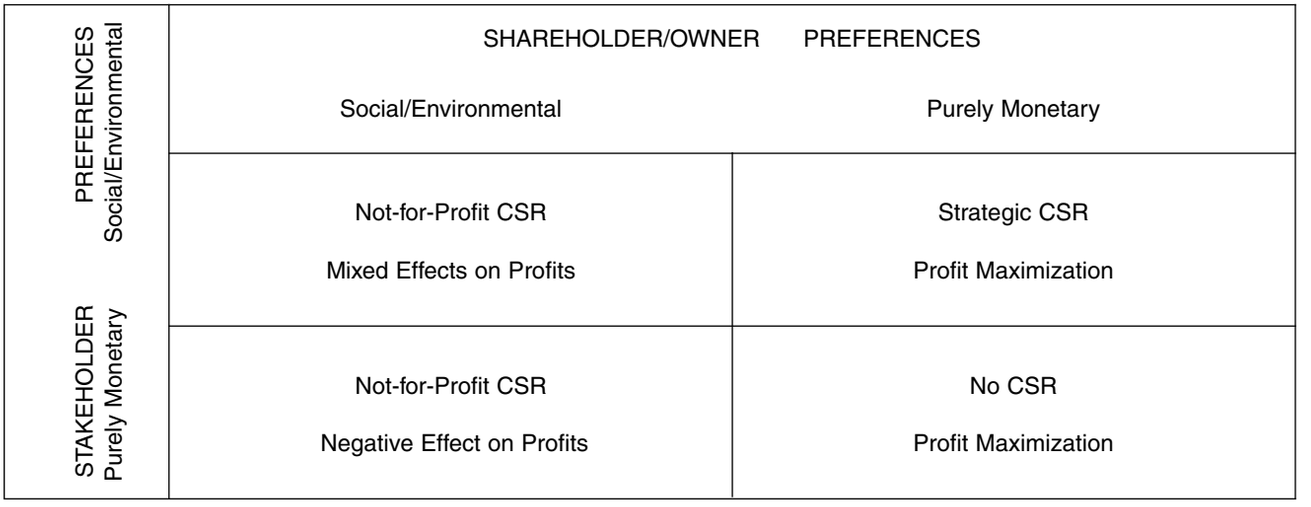


**Figure 77.1** Typology of CSR

SOURCE: Kitzmueller (2008).

recently entered microeconomic theory in general and the individual agent's utility function in particular. Standard motivational assumptions have been expanded, and a literature on intrinsic (nonpecuniary) aspects of motivation has emerged. As the behavioral economics literature is rather extensive, only a few selected contributions that are believed to improve the understanding of CSR will be reviewed.

Three general determinants of individual utility can be distinguished. Contributions by Benabou and Tirole (2003, 2006) as well as Besley and Ghatak (2005) identify (1) extrinsic (monetary) preferences and (2) intrinsic (nonmonetary) preferences as two main categories driving individual behavior via utility maximization. The intrinsic part of utility can be further divided into a (2.1) direct, nonmonetary component determined independently of how others perceive the action or payoff and (2.2) an indirect component determined by others' perception of respective action. (2.2) is frequently referred to as *reputation.* Assuming that individuals do derive utility from these three sources, economic theory can contribute to the analysis of strategic firm behavior such as CSR.

A first important insight is that intrinsic motivation can act as a substitute for extrinsic monetary incentives. Depending on the degree of substitutability, this affects both pricing through a potential increase in consumers' willingness to pay as well as "incentive design" in employment contracting (subject to asymmetric information). In sum, when pricing products, firms may be able to exploit intrinsic valuation of certain characteristics by charging higher prices, while salaries might be lower than usual if employees compensate this decrease in earnings by enjoying a social or environmentally friendly workplace, firm conduct, or reputation in line with their expectations and personal preferences. Relevant theoretic works include Benabou and Tirole (2006), who find that extrinsic incentives can crowd out prosocial behavior via a feedback loop to *reputational signaling concerns* (2.2 above). This concern reflects the possibility that increased monetary incentives might negatively affect the agent's utility as observers are tempted to conclude greediness rather than social responsibility when observing prosocial actions. Here the signal extraction problem arises because agents are heterogeneous in their valuation of social good and reputation, and this information is strictly private. Such considerations could influence not only employees, consumers, and private donors but also social entrepreneurs. Social entrepreneurs are individuals ready to give up profits or incur losses by setting up and running a CSR firm. (The opposite would be the private entrepreneur, who creates a firm if and only if its market value exceeds the capital required to create it.) CSR here expands the "social" individual's opportunity set to do "good" by the option to create a CSR firm. Summing up, agents are motivated by a mixture of extrinsic and intrinsic factors, and therefore potential nonintended effects due to crowding between extrinsic and intrinsic motivators should be taken into account when designing optimal incentives (salaries, bonus payments, taxes, etc.).

So stakeholders demand CSR in line with their intrinsic motivation, and the key question that follows, this time with respect to alternative private ways of doing social good, asks why this *corporate channel* of fulfilling one's need to do public good is used at all if there are alternatives such as direct social contribution (e.g., charitable donations or voluntary community work). From a welfare perspective, there should be some comparative advantage of CSR, something that makes it more efficient than other options. In an important paper, Andreoni (1989) compares different ways to contribute to social good and asks whether they constitute perfect or rather imperfect substitutes. Although the initial version compares public and private provision of public goods, the same analysis can be extended to compare various ways of private provision such as corporate and individual social responsibility. The answer then is straightforward. If *warm glow* effects (Andreoni, 1990) of individual (direct) altruistic giving exist, then investment into a CSR firm, government provision of public goods, and direct donations will be imperfect substitutes in utility and therefore imperfectly crowd out each other. In other words, a socially responsible consumer might not derive the same utility from buying an ethical product and from donating (the same amount of) money to charitable organizations directly. However, this analysis is unable to explain in more detail why individuals allocate a share of their *endowment to do social good* to CSR. A reasonable conjecture might be that people must or want to consume certain private goods but derive intrinsic disutility (e.g., bad conscience) from being connected to any socially stigmatized, unethical behavior or direct negative externality related to their purchase, seller, and/or use of the good or service. Furthermore, one should notice that social or environmental goods do not always directly or physically affect consumers but rather are feeding through to individual utility indirectly via intrinsic, reputational concerns (e.g., Nike's connection to child labor in Asian sweatshops). However, these conjectures have yet to be sufficiently tested empirically.

A final economic puzzle worth thinking about is the one of causality between preferences and CSR—that is, opposite to the above assumed causality from preferences to firm behavior, CSR often has been connected with advertisement or public relations of firms, thereby suggesting that CSR eventually could determine or change preferences and ultimately individual behavior over time. While the management literature has approached these issues via the concept of corporate social marketing (Kotler & Lee, 2004), economists have been more cautious when it comes to *endogenous preferences.* As far as preference formation is concerned, Becker (1993) concluded that "attitudes and values of adults are . . . influenced by their childhood experiences." Bowles (1998) builds the

bridge from Becker's "family environment" to markets and other economic institutions influencing the evolution of values, preferences, and motivations. Simon (1991) was among the first to argue that agency problems may be best overcome by attempting to change and ideally align preferences of workers and principals. The following real-world example shall illustrate the key issue here. Empirical evidence from the 1991 General Social Survey (outlined in Akerlof & Kranton, 2005, p. 22) suggests that workers strongly identify with their organization (i.e., employer's preferences). In theory, this finding can be a result of matching (selection), reducing cognitive dissonance (psychology), or induced convergence of preferences (endogenous preferences). Given these alternatives, CSR could be either interpreted as a signal leading to matching of firms and individuals with similar preferences or alternatively used to align agents' preferences over time. While the latter suggestion lacks theoretic or empirical treatment, the former potential matching role of CSR has been analyzed and will be outlined below.

It can be seen that a lot of open questions need to be answered when it comes to the mechanics of intrinsic motivation and social preferences within the human mind. Hence, further discussion of CSR focuses on strategic interaction between firms and stakeholders and treats the existence of intrinsic preferences as exogenously given.

## Six Strategies Behind Strategic CSR

Six relevant economic frameworks within which strategic CSR can arise are discussed and linked to empirical evidence at hand: (1) labor markets, (2) product markets, (3) financial markets, (4) private activism, (5) public policy, and (6) isomorphism.

### Labor Economics of CSR

CSR may alter classical labor market outcomes. Bowles, Gintis, and Osborne (2001) address the role of preferences in an employer-employee (principal-agent) relationship. The main idea is that employees might have general preferences such as sense of personal efficacy that are able to compensate for monetary incentives and therefore allow the employer to induce effort at lower monetary cost. Besley and Ghatak (2005) establish a theoretic framework to analyze the interaction of monetary and nonmonetary incentives in labor contracts within the nonprofit sector. They refer to not-for-profit organizations as being mission oriented and conjecture that such organizations (e.g., hospitals or universities) frequently are staffed by intrinsically motivated agents (think of a doctor or professor who has a nonpecuniary interest in the hospital's or university's success, i.e., saving lives or educating students). The main conclusion from their moral hazard model with heterogeneous principals and agents is that pecuniary, extrinsic incentives such as bonus payments and the agents' intrinsic motivation can act as substitutes. In other words, a match between a mission-oriented principal and an intrinsically motivated agent reduces agency costs and shirking (i.e., putting lower effort when the principal cannot observe the work effort given by the agent but only the outcome of the work) and allows for lower incentive payment. As today many firms adopt missions (such as CSR activities) in their quest to maximize profits, this analysis may directly carry over to the private sector.

As opposed to Friedman's (1970) concern that CSR is a general form of moral hazard (here moral hazard refers to managers or employees not acting in the best interest of shareholders or firm owners), Brekke and Nyborg (2004), based on Brekke, Kverndokk, and Nyborg (2003), show that CSR can actually reduce moral hazard in the labor market context. More precisely, CSR can serve as a screening device for firms that want to attract morally motivated agents. This view on CSR as a device to attract workers willing to act in the best interest of the principal is again based on the same substitutability of motivation due to CSR and related firm characteristics valued by the employees and high-powered incentives such as bonus payments.

Another labor market context that involves CSR and corporate governance is explored by Cespa and Cestone (2007). They conjecture that inefficient managers can and will use CSR (i.e., the execution of stakeholder protection and relations) as an effective entrenchment strategy to protect their jobs. CEOs and managers engage in CSR behavior in face of a takeover or replacement threat in order to then use such "personal" ties with stakeholders to bolster their positions within the firm (in other words, such managers establish themselves as key nodes linking the firm with strategic stakeholders, thereby gaining value independent of their true managerial capacity and performance). This discussion of the effect of corporate governance institutions on firm value leads to the conclusion that institutionalized stakeholder relations (as opposed to managers' discretion) close this "insurance" channel for inefficient managers and increase managerial turnover and firm value. This finding clearly provides a rationale for the existence of special institutions such as ethical indices or social auditors and increased interaction between social activists and institutional shareholders in general. A similar approach to CSR is taken by Baron (2008), who links managerial incentives and ability with the existence of socially responsible consumers. He concludes that, given that consumers value CSR, when times are good, a positive correlation emerges between CSR and financial performance via the fact that high-ability managers tend to contribute more to CSR than low-ability ones, and the level of both, CSR and profits, is increasing in managers' ability. In bad times, however, shareholders are not supporting social expenditure (for profits) anymore, high-ability managers become less likely to spend money on CSR as compared to

low-ability ones, and the correlation between CSR and profits becomes negative. Baron's work gives a first idea of the importance of consumer preferences in the determination of CSR efforts, which will be the subject of the following subsection.

## CSR and Product Markets (Socially Responsible Consumption)

With regard to whether consumers really care about CSR, there is substantial empirical evidence supporting this assumption. Consumer surveys such as the Millennium Poll on Corporate Social Responsibility (Environics International Ltd., 1999) or MORI (http://www.ipsos-mori.com/about/index.shtml) reveal that consumers' assessment of firms and products as well as their final consumption decisions and willingness to pay depend on firms' CSR records. In this respect, Trudel and Cotte (2009) find the equivalent to loss aversion in consumers' willingness to pay for ethical products. According to their findings, consumers are willing to pay a premium for ethical products and buy unethical goods at a comparatively steeper discount. So there exists a channel from preferences and demand to CSR and/or vice versa.

Consumer preferences may translate into demand for CSR and alter the competitive environment of firms as CSR can either act as product differentiation or even trigger competition with respect to the level of CSR itself. Bagnoli and Watts (2003) analyze competitive product markets with homogeneous, socially responsible consumers and find that CSR emerges as a by-product and at levels that vary inversely with the degree of competitiveness in the private goods market (competitiveness is reflected through both number of firms and firm entry). Bertrand (price) as opposed to Cournot (quantity) competition forces firms to reduce markups and hence limits their ability to use profits to increase CSR. This leads to reduced competitiveness in terms of product differentiation via CSR and hence to reduced overall CSR activity. In sum, there exists a trade-off between efficient provision of the private good and public good (i.e., Bertrand competition entails lower prices and lower levels of CSR than Cournot competition).

A more general framework is provided by Besley and Ghatak (2007), who find that Bertrand (price) competition in markets with heterogeneous demand for CSR leads to zero profits—that is, prices equal marginal costs and second best (suboptimal) levels of public good provision equivalent to results obtained in models of private provision (e.g., Bergstrom et al., 1986). Their analysis further allows validation of a whole array of standard results from the screening and public goods literature, among which, (a) the maximum sustainable level of CSR (under imperfect monitoring by consumers) is achieved when the firms' incentive compatibility constraint binds—that is, at such a public good level the profits from doing CSR and charging a price premium are equivalent to profits from not producing the public good and still charging a price premium (cheating), given any probability (between 0 and 1) of being caught cheating and severely punished. (b) An exogenous increase of public good supply (e.g., by a government through regulation or direct production) perfectly crowds out competitive provision of CSR. (c) In the absence of government failure, governments are able to implement the first best Lindahl-Samuelson level of public good (i.e., the Pareto optimal amount, which is clearly above the levels markets can provide via CSR). (d) However, when governments fail (e.g., due to corruption, capture, relative production inefficiencies, or distributional bias), CSR might generate a Pareto improvement vis-à-vis no production or government production of public good, while CSR and provision by nonprofits (e.g., NGOs) are identical unless one or the other has a technological (cost) advantage in producing the public good. (e) Finally, a small uniform regulation in the form of a minimum standard (on public good levels) would leave the level of CSR unchanged and redistribute contributions from social to neutral consumers, while large regulatory intervention can raise supply of the public good to or above its first best level given that neutral consumers are willing to pay higher (than marginal cost) prices for the private good. These results highly depend on respective consumer preferences and their related willingness to pay for the private and public good (CSR) characteristics of the consumption good.

Arora and Gangopadhyay (1995) model CSR as firm self-regulation (i.e., voluntary overcompliance with environmental regulation) and assume that although consumers all value environmental quality, they vary in their willingness to pay a price premium for CSR, which is positively dependent on their income levels. Firms differentiate by catering to different sets of consumers; here choice of *green* technology acts as product positioning similar to the choice of product quality, and CSR is positively correlated with the income levels of either all consumer segments or the lowest income segment. If a minimum standard is imposed into a duopoly, it will actually bind on the less green firm while the other firm will overmeet the standard. CSR subsidies can have the same effect as standards, while ad valorem taxes (i.e., taxes on profits) always reduce output and CSR efforts by all firms.

An example of empirical work in this subfield is provided by Siegel and Vitaliano (2007), who test and confirm the hypothesis that firms selling experience or credence goods—that is, the good's quality can only be observed after consumption (by experience, e.g., a movie) or is never fully revealed (credence, e.g., medical treatment or education)—are more likely to be socially responsible than firms selling search goods (i.e., goods where characteristics such as quality are easily verified ex ante). This lends support to the conjecture that consumers consider CSR as a signal about attributes and general quality when product characteristics are difficult to observe. From the

firm perspective, CSR then can be used to differentiate a product, advertise it, and build brand loyalty. The advertising dimension of CSR is especially strong when social efforts are unrelated to business conduct. In Navarro (1988), corporate donations to charity are identified as advertisement, and CSR is meant to transmit a positive signal about firm quality/type. However, according to Becker-Olsen and Hill (2005), this signal might not necessarily be positive, as consumers are able to identify low-fit CSR as advertisement and tend to negatively perceive such CSR efforts as greediness of firms rather than genuine interest in social or environmental concerns.

## CSR and Financial Markets (Socially Responsible Investment)

Investors also care about CSR, and firms competing for equity investment in stock markets will have to take that into account. Geczy, Stambaugh, and Levin (2005) put forward strong evidence of the increasing importance of CSR in financial markets. A new form of investment, so-called *socially responsible investment* (SRI), has come into being. SRI is defined by the Social Investment Forum (SIF, 2009) as *an investment process that considers the social and environmental consequences of investments, both positive and negative, within the context of rigorous financial analysis.* Social investors today include both private and institutional ones. More precisely, the U.S. Social Investment Forum (figures are taken from SIF, 2006) reports 10.8% of total investment under professional management in 2007 to be socially responsible (i.e., using one or more of the three core socially responsible investing strategies: screening, shareholder advocacy, and community investing). In Europe, the European Sustainable and Responsible Investment Forum (EuroSIF) identifies 336 billion euros in assets to be SRI. The trend points upward in most financial markets (e.g., in the United States, where SRI assets grew 4% faster than total assets and more than 258% in absolute terms between 1995 and 2005).

Recalling the typology of CSR (Figure 77.1), we know that investors either have or do not have social preferences. Neutral investors just have their monetary return on investment in mind and hence just care about firm profits. It follows that such investors will use SRI as an investment strategy only if SRI actually translates into higher returns on investment. So, SRI by neutral investors signals a comparative advantage in corporate financial performance (CFP). This conjecture and the related question of correlation and causality have attracted a lot of attention in the scarce empirical literature on CSR. A comprehensive survey is provided by Margolis and Walsh (2003). Taking into account 127 published empirical studies between 1972 and 2002, they find that a majority of these studies exhibit a statistically significant and positive correlation between CSR and CFP in both directions (i.e., causality is running from CFP to CSR and vice versa). However, there exist sampling problems, concerns about the validity of CSR and CFP measures and instruments, omitted variable bias, and the ultimate (and still unanswered) question of causality between CSR and CFP. A first attempt to address inconsistency and misspecification is the work by McWilliams and Siegel (2000), who regress firm financial performance on CSR and control for R&D investment. It follows that the upwards bias of the financial impact of CSR disappears and a neutral correlation emerges. In sum, further studies will have to clarify whether neutral investors should put their money into SRI and the underlying CSR effort qualifies as *strategic.*

Alternatively, SRI can be a way for social investors to enforce their preferences through a demand channel similar to the one consumers use. The group of social investors, however, can again be heterogeneous in the sense that there might be those for whom corporate giving is a close substitute for personal giving and those for whom it is a poor substitute (Baron, 2005). Small and Zivin (2005) enrich this setup by deriving a Modigliani-Miller theory of CSR, where the fraction of investors that prefers corporate philanthropy over private charitable giving drives CSR by firms attempting to maximize their valuation. A share constitutes a charity investment bundle matching social and monetary preferences of investors with those of the firm's management. The main conclusion is that if all investors consider CSR and private charity as perfect substitutes, share prices and the aggregate level of philanthropy are unaffected by CSR. If they are imperfect substitutes and a sufficiently large fraction of investors prefers CSR over private charity (e.g., to avoid corporate taxation), a strictly positive level of CSR maximizes share prices and hence the value of a corporation.

## CSR and Private Politics (Social Activism)

The existence and impact of social or environmental activists is intimately related with information asymmetries between companies and the outside world. The rationale of *social activism* is that the threat of negative publicity (revelation of negative information) due to actions by an unsatisfied activist motivates CSR. As soon as the activist is credible and has the ability to damage a firm's reputation or cause substantial costs to the firm, the existence of such an activist is sufficient to integrate CSR as part of corporate strategy. The logic is comparable to the one of "hedging" against future risk in financial markets, but here the firm insures itself against a potential campaign by an activist. Baron (2001) explicitly adds this threat by an activist, who is empowered with considerable support by the public, to the set of motivations for strategic CSR. CSR is referred to as corporate redistribution to social causes motivated by (1) profit maximization, (2) altruism, or (3) threats by an activist. However, it can be argued that the existence of activism qualifies CSR as an integral part of profit maximization (i.e., motivation 3 fuses into 1).

The main insights from the analysis of CSR and social activism can be summarized as follows: First, CSR and private politics entail a direct cost effect depending on the competitive environment (i.e., the degree of competition is

positively correlated with the power of an activist boycott and strengthens the ex ante bargaining position of the activist). On the other hand, CSR can have a strategic effect that alters the competitive position of a firm. What is meant here is that CSR can act as product differentiation (lower competition), take the wind out of the sails of any potential activist, and reduce the likelihood of being targeted in the future. This result roots in the assumption that the activist also acts strategically and chooses "projects" that promise to be successful (i.e., weaker firms are easier targets). Finally, the existence of spillover effects from one firm to other firms or even the whole industry can act as an amplifier to activist power, on one hand, and motivation for concerted nonmarket action by firms in the same industry, on the other (e.g., voluntary industry standards).

Baron (2009) assumes that citizens prefer not-for-profit (morally motivated) over strategic CSR. If signaling is possible, morally motivated firms achieve a reputational advantage, and social pressure will be directed toward strategic firms. If citizens are not distinguishing, morally motivated firms are more likely targeted as they are "softer" targets in the sense that an activist is more likely to reach a "favorable" (i.e., successful from activist objective's point of view) bargain with such a firm. However, the distinction between strategic and not-for-profit CSR can be extremely difficult, subtle, and based on perception rather than facts. Recent work by marketing scholars lends support to this proposition. Becker-Olsen and Hill (2005) find that consumers form their beliefs about CSR based on perceived fit and timing of related efforts (i.e., a high fit between CSR and the firm's business area as well as proactive rather than reactive social initiatives tend to align consumers' beliefs, attitudes, and intentions with those of the strategic firm). Finally, if activists differ in ability, Baron shows that high-quality activists attract greater contributions and then are more likely to identify and target strategic firms, while the opposite holds for low-quality activists.

### CSR and Public Politics (Regulation)

CSR is defined as corporate social or environmental effort beyond legal requirements. Then how can public politics and laws actually stimulate CSR? This time, it is the threat of future laws and regulations and the adjustment costs and competitive disadvantage they could entail that act as an incentive to hedge against such an event and build a strategic "buffer zone" via CSR. Again, by doing CSR, firms not only are "safe" in the event of regulation but also might discourage government intervention. This last point has been addressed under the label of *crowding out.* The analysis here focuses on whether market provision of public goods and public provision are substitutes or complements and how they might interact (Bergstrom et al., 1986). From a policy perspective, it seems to be important not only to understand interaction between public provision and CSR but also to consider CSR itself as a potential target for novel policies aiming to stimulate corporate provision of public goods.

Calveras et al. (2007) study the interplay between activism, regulation, and CSR and find that private (activism) and public (regulatory) politics are imperfect substitutes. It is emphasized that when society free rides on a small group of activist consumers, loose formal regulation (voted for by the majority of nonactivists) might lead to an inefficiently high externality level, where activist consumers bear the related cost via high prices for socially responsible goods. This conclusion draws attention to another relevant correlation—namely, between regulation and political orientation. Consumers are also voters, and not only firms but also governments want to signal their type (i.e., whether they value environmental or social public goods). As governments signal their future intentions and policy stances through legislation or regulation and firms through CSR, the potential competition and related interaction between regulation and CSR constitute an important subject of further investigation.

Empirically, Kagan, Gunningham, and Thornton (2003) address the effect of regulation on corporate environmental behavior. They find that regulation cannot fully explain differences in environmental performance across firms. However, "social license" pressures (induced by local communities and activists) as well as different corporate environmental management styles significantly add explanatory power. In sum, regulation matters, but variation in CSR is also subject to the antagonism between social pressure and economic feasibility.

### Isomorphism

Here, the incentive to do CSR roots in isomorphic pressures within geographic communities or functional entities such as industries. Community isomorphism refers to *the degree of conformity of corporate social performance in focus, form, and level within a community.* It is the institutional environment and commonly (locally) accepted norms, views, and values that might discipline firms into certain social behavior. Institutional factors that are potentially shaping the nature and level of CSR in a community include cultural-cognitive forces, social-normative factors, and regulative factors. Marquis, Glynn, and Davis (2007) use an institutional theoretic setting and identify community isomorphism as a potential explanatory variable for empirical observations concerning CSR. Isomorphic pressures may also arise within industries and may lead to industry-wide self-regulatory activities.

## Conclusion

From an economic point of view, a fundamental understanding of CSR is emerging. Based on a new set of social or environmental stakeholder preferences, CSR can be fully consistent with a profit- and/or shareholder value-maximizing corporate strategy. It qualifies as strategic behavior if consumers, investors, or employees have relevant social or environmental preferences and if these

preferences translate into action with direct or indirect monetary effects for the firm. Direct consequences include the firm's ability to charge price premia on CSR comparable to premia on product quality, as well as the potential lowering of wages for "motivated" employees, who substitute the utility they gain from working for and within a responsible firm/environment for monetary losses due to lower salaries. Firms' profits may be indirectly affected by CSR in the sense that CSR can help avoid competitive disadvantages or reputation loss arising in situations where stakeholder action (consumption or activism) depends on social or environmental corporate conduct. Empirical evidence lends support to most of these incentives for strategic CSR; however, rigorous statistical analysis is still in an infant state and subject to various problems, including measurement error, endogeneity, and misspecification.

*Author's Note:* Portions of this chapter appeared in different form in Kitzmueller (2008).

## References and Further Readings


Akerlof, G. E., & Kranton R. E. (2005). Identity and the economics of organization. *Journal of Economic Perspectives, 19,* 9–32.

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy, 97,* 1447–1458.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal, 100,* 464–477.

Arora, S., & Gangopadhyay, S. (1995). Toward a theoretical model of voluntary overcompliance. *Journal of Economic Behavior and Organization, 28,* 289–309.

Bagnoli, M., & Watts, S. G. (2003). Selling to socially responsible consumers: Competition and the private provision of public goods. *Journal of Economics and Management Strategy, 12,* 419–445.

Baron, D. P. (2001). Private politics, corporate social responsibility, and integrated strategy. *Journal of Economics and Management Strategy, 10,* 7–45.

Baron, D. P. (2005). *Corporate social responsibility and social entrepreneurship* (Stanford GSB Research Paper No. 1916). Stanford, CA: Stanford University Press.

Baron, D. P. (2008). Managerial contracting and CSR. *Journal of Public Economics, 92*(1–2), 268–288.

Baron, D. P. (2009). A positive theory of moral management, social pressure and corporate social performance. *Journal of Economics and Management Strategy, 18,* 7–43.

Becker, G. S. (1993). Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy, 101,* 385–409.

Becker-Olsen, K. L., & Hill, R. P. (2005). The impact of perceived corporate social responsibility on consumer behavior. *Journal of Business Research, 59,* 46–53.

Benabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies, 70,* 489–520.

Benabou, R., & Tirole, J. (2006). Incentives and pro-social behavior. *American Economic Review, 96,* 1652–1678.

Bergstrom, T., Blume, L., & Varian, H. R. (1986). On the private provision of public goods. *Journal of Public Economics, 29,* 25–49.

Besley, T., & Ghatak, M. (2001). Government versus private ownership of public goods. *Quarterly Journal of Economics, 116,* 1343–1372.

Besley, T., & Ghatak, M. (2005). Competition and incentives with motivated agents. *American Economic Review, 95,* 616–636.

Besley, T., & Ghatak, M. (2007). Retailing public goods: The economics of corporate social responsibility. *Journal of Public Economics, 91,* 1645–1663.

Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature, 36,* 75–111.

Bowles, S., Gintis, H., & Osborne, M. (2001). Incentive-enhancing preferences: Personality, behavior and earnings. *American Economic Review, 91*(2), 155–158.

Brekke, K. A., Kverndokk, S., & Nyborg, K. (2003). An economic model of moral motivation. *Journal of Public Economics, 87,* 1967–1983.

Brekke, K. A., & Nyborg, K. (2004). *Moral hazard and moral motivation: Corporate social responsibility as labor market screening* (Memorandum N025). Oslo, Norway: Department of Economics, University of Oslo.

Buchanan J. M. (1999). *The demand and supply of public goods.* Indianapolis, IN: Liberty Fund. Retrieved from http://www.econlib.org/Library/Buchanan/buchCv5c1.html

Calveras, A., Ganuza, J. J., & Llobet, G. (2007). Regulation, corporate social responsibility, and activism. *Journal of Economics and Management Strategy, 16,* 719–740.

Cespa, G., & Cestone, G. (2007). Corporate social responsibility and managerial entrenchment. *Journal of Economics and Management Strategy, 16,* 741–771.

Economist Intelligence Unit. (2005, January). *The importance of corporate responsibility* (White paper). Retrieved from http://graphics.eiu.com/files/ad_pdfs/eiuOracle_CorporateResponsibility_WP.pdf

Environics International Ltd. (1999, September). Millennium poll on corporate responsibility, in cooperation with The Prince of Wales Trust.

European Commission. (2009). *Corporate social responsibility (CSR).* Available at http://ec.europa.eu/enterprise/policies/sustainable-business

Fleishman-Hillard & the National Consumers League. (2005). *Rethinking corporate social responsibility.* Retrieved from http://www.csrresults.com/FINAL_Full_Report.pdf

Friedman, M. (1970, September 13). The social responsibility of business is to increase its profits. *The New York Times,* pp. 32–33, 122, 126.

Geczy, C., Stambaugh, R., & Levin, D. (2005). *Investing in socially responsible mutual funds* (Working paper). Philadelphia: Wharton School of Finance.

Kagan, R. A., Gunningham, N., & Thornton, D. (2003). Explaining corporate environmental performance: How does regulation matter? *Law and Society Review, 37,* 51–90.

Kitzmueller, M. (2008). *Economics and corporate social responsibility* (EUI ECO Working Paper 2008–37). Available at http://cadmus.eui.eu

Kotchen, M. J. (2006). Green markets and private provision of public goods. *Journal of Political Economy, 114,* 816–845.

Kotler, P., & Lee, N. (2004, Spring). Best of breed. *Stanford Social Innovation Review,* pp. 14–23.

Margolis, J. D., & Walsh, J. (2003). Misery loves companies: Rethinking social initiatives by business. *Administrative Science Quarterly, 48,* 268–305.

Marquis, C., Glynn, M. A., & Davis, G. F. (2007). Community isomorphism and corporate social action. *Academy of Management Review, 32,* 925–945.

McWilliams, A., & Siegel, D. S. (2000). Corporate social responsibility and financial performance: Correlation or misspecification? *Strategic Management Journal, 21,* 603–609.

McWilliams, A., & Siegel, D. S. (2001). Corporate social responsibility: A theory of the firm perspective. *Academy of Management Review, 26,* 117–127.

Navarro, P. (1988). Why do corporations give to charity? *Journal of Business, 61,* 65–93.

Reinhardt, F. L., Stavins, R. N., & Vietor, R. H. K. (2008). *Corporate social responsibility through an economic lens* (NBER Working Paper 13989). Cambridge, MA: National Bureau of Economic Research.

Rose-Ackerman, S. (1996). Altruism, nonprofits and economic theory. *Journal of Economic Literature, 34,* 701–728.

Scherer, A. G., & Palazzo, G. (2008). Globalization and corporate social responsibility. In A. Crane, A. McWilliams, D. Matten, J. Moon, & D. Siegel (Eds.), *Oxford handbook of corporate social responsibility* (pp. 413–431). Oxford, UK: Oxford University Press.

Siegel, D. S., & Vitaliano, D. F. (2007). An empirical analysis of the strategic use of corporate social responsibility. *Journal of Economics and Management Strategy, 16,* 773–792.

Simon, H. A. (1991). Organizations and markets. *Journal of Economic Perspectives, 5*(2), 25–44.

Small, A., & Zivin, J. (2005). A Modigliani-Miller theory of altruistic corporate social responsibility. *Topics in Economic Analysis and Policy, 5*(1), Article 10.

Social Investment Forum. (2006). *2005 report on socially responsible investing trends in the United States.* Retrieved from http://www.socialinvest.org/pdf/research/Trends/2005%20Trends%20Report.pdf

Social Investment Forum. (2009). *The mission in the marketplace: How responsible investing can strengthen the fiduciary oversight of foundation endowments and enhance philanthropic missions.* Available at http://www.socialinvest.org/resources/pubs

Stiglitz, J. E. (1993). Post Walrasian and post Marxian economics. *Journal of Economic Perspectives, 7,* 109–114.

Stiglitz, J. E. (2002). Information and the change in the paradigm in economics. *American Economic Review, 92,* 460–501.

Trudel, R., & Cotte, J. (2009). Is it really worth it? Consumer response to ethical and unethical practices. *MIT/Sloan Management Review, 50*(2), 61–68.

Vogel, D. (2005). *The market for virtue: The potential and limits of corporate social responsibility.* Washington, DC: Brookings Institution Press.

World Bank. (n.d.). *Financial and private sector development.* Retrieved from http://www.ifc.org/ifcext/economics.nsf/content/csr-intropage

# 78

# Political Economy of Violence

Shawn Humphrey
*University of Mary Washington*

Long-term economic prosperity requires the existence of secure property rights. A resource, in the possession of an individual, can be characterized as secure property if that individual can exclude others from exploiting its use. So, as the fundamental instrument of exclusion, who should have recourse to violence? In a world where violence is a recognized margin of competitive adjustment available to all community members, a world of *unorganized* violence, an individual's exclusive right to use a resource is determined in large part by his or her ability to use naked violence to protect that resource from others. This fact creates an incentive for all community members to allocate scarce effort away from production and toward protection and/or predating on others—both of which undermine a community's potential for prosperity.

A solution to this inefficient outcome is to structure a world of *organized* violence, a world where violence is managed and access to it is limited to a subset of the community. In other words, community members can choose to structure a political economy. One such example is the state. In its simplest form, the state consists of a third party (e.g., a prince, king, queen, prime minister, and/or president) and constituents. The fundamental exchange undergirding the state involves constituents surrendering their autonomy, forgoing their rightful recourse to violence, and providing some sort of compensation (e.g., tax revenue) to the third party in exchange for secure property rights. Decisions regarding the projection of violence are reserved for the third party. Structuring a political economy also includes deciding on the number, character, and composition of those that the third party can call upon to violently enforce property rights (e.g., citizens militias, police, military and/or paramilitary forces). However, when deciding how to organize violence, the members of any community confront a dilemma. Centralizing too much violence potential in too few hands can be an inefficient alternative to a world of *unorganized* violence. The power disparity between the third party and its constituents can create an incentive for the third party to predate on its constituents and consequently destabilize its own property rights regime. Constituents can confront this challenge by dissipating the third party's violence potential. This can be accomplished by choosing a more decentralized organization of violence—for example, one in which access to violence is granted to more individuals other than the third party and his or her forces (e.g., a citizens militia). However, this organization can weaken the third party and invite external and internal rivals to challenge the established order, thereby also destabilizing property rights. In other words, economic prosperity requires the presence of a third party that is powerful enough to establish and enforce property rights but not so powerful that its presence destabilizes these rights. This tension is known as the credible commitment dilemma. Successfully managing this tension is pivotal to secure property rights and hence economic prosperity.

## Theory

### Violence and Economic Prosperity

The fundamental purpose of this chapter is to explore the role that violence plays in the determination of a community's economic prosperity. We begin by acknowledging the fact that we live in a world of scarcity. There are simply not enough resources available to satisfy every want. We also reside in a world populated by individuals with a multiplicity of desires, wants, and preferences. Alternative uses are envisioned for some, if not most, and maybe all

scarce resources. Competition—two or more individuals placing simultaneous demands upon the same resource—is our constant companion. In general, if one individual currently possesses something that another desires to allocate toward an alternative use, there is one of two ways that such a conflict of interest can be resolved. The individual who does not currently possess the resource can trade away something he or she does possess in exchange for that which he or she desires, or the individual can forcibly take it. Whether this interaction results in attempted trading or taking is contingent on whether or not possession, by the individual who currently holds the resource, translates into exclusive ownership. In other words, does this individual have secure property rights over this resource? Property rights, in a world of scarcity and competition, are only as exclusive as others around you allow them to be. Those around you have to decide whether or not to respect your rights. Respect can be earned by your ability to personally exclude the encroachments of others and/or bestowed by your community. The latter is most conducive to a voluntary value-creating rearrangement of resources and therefore economic prosperity.

Among neighboring individuals, violence has been and continues to be a fundamental way of deciding issues of ownership—this is mine and this is yours. Within the confines of this chapter, *violence* is defined as the use or threatened use of scarce resources for the sake of imposing physical costs on others. We will define *unorganized violence* as a context in which violence is a widely recognized and practiced method of establishing exclusivity. As you may expect, unorganized violence is inimical to economic prosperity. It can lead to the absolute destruction of resources and productive assets (e.g., the killing and maiming of people and livestock and the destruction of infrastructure—including roads, bridges, hospitals, and schools). It can also create an incentive for individuals to redirect their efforts from otherwise growth-augmenting endeavors—specifically taking the steps to enhance their individual productivity.

Productivity—how much output an individual can produce in one hour of his or her time—is the key to economic prosperity. An individual's productivity is a function of the degree of specialization in his or her community and his or her access to capital (both physical and human) and technological innovations. Choosing to specialize, gaining access to capital, and taking advantage of technological innovations are investments that the members of a community must be willing and able to take. Investments, however, entail a trade-off. The individual must forgo some level of current consumption and other alternative uses of his or her time in exchange for enhanced future productivity and returns. For some extended period of time, this person is tying up today's scarce resources in the pursuit of future returns. Returns that lie in the future are returns that can never be guaranteed. Indeed, the longer that resources are tied up in a particular investment, the greater the risk of conditions changing and thereby threatening his or her returns. Every investment, necessarily, is a risky action that not everyone will find worthwhile to take.

Ubiquitous violence only makes making investments more risky. For example, knowing that everyone else, like you, has recourse to violence, you confront a fundamental trade-off when attempting to arrive at an income-maximizing allocation of effort. Namely, in the context of unorganized violence, your exclusivity over resources is determined in large part by your comparative advantage in violence. Necessarily, every unit of effort invested in enhancing your productivity raises your opportunity cost of allocating effort toward violence. Enhancing your productivity dissipates your comparative advantage in violence. Indeed, your enhanced income-producing potential combined with your diminished willingness to enforce an exclusive claim to that income (it has simply become more costly to exclude) makes you a more lucrative target. Your investment in the future motivates those around you to predate upon you. Looking forward to this possible outcome, you may choose against this act. Indeed, you may enjoy relatively more secure property rights—albeit over limited property and diminished income streams—by being less productive. Choosing relative poverty may insure you against being victimized.

Beyond adversely influencing the level of investment in the economy, unorganized violence induces a number of other growth-discouraging dynamics. For example, you may enjoy a higher standard of living by simply protecting what you already have from your neighbors or by predating upon them. Necessarily, you may have an incentive to invest scarce time and effort into enhancing your protective and/or predatory potential. These investments include the development of your potential to impose physical costs on others by investigating new techniques and/or training in traditional techniques of violence. These investments can also include the steps taken to diminish the potential of competitors to impose costs on you by building fortifications, training dogs, and finding hiding places for your wealth. Protective measures can also include shying away from interacting with others—which is not conducive for the creation markets.

A market exists whenever a buyer and seller of a particular good or service conclude that their interaction is mutually beneficial. The extent of a market for a particular good or service is determined by the number of mutually beneficial exchanges that occur between buyers and sellers. Necessarily, any force that diminishes these mutual benefits—the gains from trade—threatens the existence and extent of markets. A number of forces can influence the gains from trade. We are concerned with transaction costs—the cost of engaging in exchange—which include search, bargaining, and enforcement costs. Our focus is on enforcement costs. In particular, before the conclusion of an exchange, a buyer may ask, "Will I get what I paid for?" Conversely, the seller may ask, "Will I get paid?" An

exchange environment that is characterized by mutual trust and hence low enforcement costs allows both to answer the aforementioned questions affirmatively and capture the gains from trade. However, in the context of unorganized violence, where mutual trust is lacking, enforcement costs will be significant. Consequently, in this exchange environment, the gains from trade will fall, fewer and fewer exchanges will be concluded, and markets will either shrink or fail to exist.

This is important because the returns that attend all three aforementioned productivity-enhancing investments are positively associated with the existence and extent of markets. If markets do not exist, then there is no opportunity to make investments. Even if markets exist, if their extent is limited, then the returns for these investments may be too low to make them worthwhile.

Unorganized violence also creates an incentive for you to manage your possessions and resources in such a way so as to maximize their liquidity. You can accomplish the latter by not tying up your resources in investments whose benefits are delayed. Your neighbors confront the very same trade-offs. In consequence, market size will be reduced, and what property is possessed will remain portable to enhance its liquidity. Incentives are such that it may be individually rational to forgo specialization, productivity-enhancing pursuits, and potentially mutually advantageous exchanges with others. Either way no one individual in the community is taking the costly and risky steps toward becoming a productivity-enhancing specialist short-circuiting the emergence of markets and disabling the fundamental engine of economic prosperity.

## Why the State Is Necessary for Economic Prosperity

Social order, the opposite of unorganized violence, is widely recognized as a necessary condition for long-term economic prosperity. Order exists when community members find it worthwhile, given their expectations about the actions of others, to respect and defend the rules that delineate claims of ownership. By adhering to a set of rules and agreeing not to predate upon each other, members can reduce the amount of resources allocated toward protective purposes. By agreeing to punish predators in their midst, the returns to predation fall, hence discouraging this activity. Moreover, by working together, they may be able to realize a higher level of exclusivity at lower cost. All possibilities free up scarce resources for productivity-enhancing pursuits and allow for the value-creating rearrangement of resources to take place through trade as opposed to violent struggles. Yet, how does a community of rational self-interested individuals realize social order?

As opposed to the status quo (unorganized violence), social order provides an opportunity for mutually beneficial interactions. Necessarily, there will be a demand for some property rights to govern the community's interactions. It is the supply, monitoring, and enforcement of these rights that are the fundamental hurdles to the emergence of order (Ostrom, 1990). The supply of a set of rules that yields social order will make the whole better off; however, the rules may not make everyone equally better off. Disagreements, necessarily, may emerge over which set of rules to put in place. Even in the case of symmetric benefits, the costs of supplying these rules may not be borne equally by all members. Moreover, supplying a new set of rules is a public good. Individual members have an incentive to secure these benefits without bearing any cost. Ignoring these difficulties, whether or not a set of rules yields benefits above and beyond those that attend unorganized violence, is contingent on whether or not those subscribing to them actually abide by them. It is here that additional difficulties arise.

One way to realize social order is by some or all individuals making up a community to choose to cooperate in a mutual defense pact, which obligates everyone to respect and defend the rights of others. Respecting the rights of others is costly. It requires one to accept the way in which ownership is defined and structured. The individual bears all these costs—the forgone benefits one could have enjoyed by simply taking what he or she deemed desirable—yet enjoys only a share of the benefits that attend the consequent stability. Similarly, when defending the rights of another, an individual bears fully the opportunity cost of the resources allocated toward this task but enjoys only a share of the benefits. In both cases, the community at large enjoys the remaining share. Respecting and defending the rights of others are public goods. Every member of the community has an incentive to free ride on the effort of others. Everyone in the group is susceptible to this temptation. Everyone in the group knows that everyone in the group is susceptible to this temptation. No one member of the mutual defense pact can truly trust that others will abide to the agreement. This heightens the temptation to renege even more—all choose to be violent. It is their dominant strategy.

This outcome can be avoided if everyone commits to respecting the rules. The idea here is that by committing to respect the rules, you can effectively alter others' expectations concerning your behavior, and by altering their expectations, you can alter the behavior of those around you. They expect you to respect and defend the rules, which in turn favorably alter their marginal net return to respecting and defending the rules. The process of committing yourself and the group may be as simple as you and the group asking each other, "Will you abide by the rules?" The question is whether or not your respective replies, "I will abide" and "We will abide," are believable. Will they alter your expectations concerning their behavior? Will you change their expectations concerning your behavior? They may not believe you. They may not be telling you the truth. Your respective replies lack credibility for the simple reason that they go against rationality.

Rationality dictates that an individual should ex ante promise to abide by the rules and ex post renege when

everyone has invested costly effort into productive pursuits. The cost of a promise is negligible. And if everyone else cooperates, you can enjoy a share of the output they produce by predating upon them. It is possible that other members of the group will attempt to monitor your actions and, if necessary, exclude you from the benefits that attend order; however, two things make it possible for your strategy to be successful. Like the supply of rules, monitoring is a public good. It will be undersupplied by the rest of the group. Moreover, even if you were caught attempting to predate, you could expect not to be punished. Enforcement is also a public good. It will not be supplied, by the group, at the efficient level. This logic holds for every member of the group. They all can enjoy the benefits of order without bearing the costs. Commitments, without effective monitoring and enforcement, lack credibility (Ostrom, 1990). Neither your expectation concerning their behavior nor their expectations concerning your behavior will change. Consequently, forward-looking individuals, expecting both services to be undersupplied, may find it worthwhile to continue to invest in protecting what they have. Indeed, it may be worthwhile to continue to encroach, take, rob, and steal. Prospective members are also cognizant of the fact that when conditions change, those who agree to be bound by the rules today may find it worthwhile to break them tomorrow. Given the aforementioned difficulties, we would not expect a collective of rational self-interested individuals to be capable of supplying order and self-regulate violence in their community. Even if one member of the group were willing to bear solely the cost of supplying a set of rules, looking forward, he or she would expect that commitments lack credibility. Consequently, any effort this individual would allocate toward structuring a set of rules would be a fruitless endeavor (Ostrom, 1990). Unorganized violence, however, does not characterize all communities (Molinero, 2000). In the following sections, we review the conditions that can yield credible and noncredible commitments and how changes in these conditions are intimately related to the emergence of the state.

Order can arise by voluntary agreement when groups are small for two reasons. With the benefits of supplying social order shared among a smaller number of individuals, the net advantages of doing so may be positive (Olson, 1965). Even in the presence of the positive externality, individuals may still be willing to respect and defend others' property rights. Smaller numbers also facilitate one's ability to supply order. Individuals in smaller communities are for the most part related to each other through blood, marriage, or both (Diamond, 1999). When a conflict does arise, the parties to the conflict will share many kin who have the ability to intervene, adjudicate, and apply pressure for a nonviolent resolution (Diamond, 1999). Moreover, the relationships among members of small communities are not characterized by one-shot interactions—they are repeated. If one member of the collective is deciding whether or not to act against the prevailing order, the lost opportunity of no longer engaging in trade with the rest of the community can be a forceful constraint (Axelrod, 1985).

The larger share of benefits enjoyed coupled with the familiarity and repeated play that characterizes small communities are favorable conditions for the emergence of order; however, they can be ephemeral. The rising living standards that accompany social order, including increased life expectancy and decreased mortality rates, translate into population growth. Population growth brings opportunities and challenges. Additional community members, with their unique preferences and abilities, can allow for the introduction of new goods and services and the lowering of transformation costs, through competition and economies of scale. Capturing them, however, is contingent on the successful conclusion of an increased number of exchanges with multiple trading partners. It is here where difficulties crop up. An increase in the number, and maybe the diversity, of exchange partners increases the complexity of the economic-exchange environment. These new exchange partners may require monitoring because having another exchange partner(s) limits the need to return again and again to a previous exchange partner. Forward-looking exchange partners recognize that it may be worthwhile for the partner in trade to renege. Consequently, more monitoring will be required, and exchanges may become more costly to transact, limiting the potential to capture the aforementioned gains that attend population growth.

It is also reasonable to assume that beyond a certain population threshold, social order begins to break down (Olson, 1965). Large numbers erode the familiarity and repeated play that characterized earlier interactions. They also diminish the share of the benefits that attend supplying order. Both forces adversely influence the net advantages of this activity. Knowing that others like you are unmotivated to supply, monitor, and enforce order, individual maximization predicts that if one's interests are better served by breaking the predominant order, then one will do just that. Necessarily, conflicts of interests become more frequent. Moreover, even if you wanted to monitor and enforce order, your ability to do so may be handicapped by the simple fact that you may no longer know both sides to a disagreement. Commitments begin to lose their credibility. Conflicts become more frequent and can escalate if either one or both parties to a dispute summon friends and family members to aid them in an armed threat or attack on the other. The community's ability to self-regulate violence begins to erode.

Unorganized violence, however, may not return. It can be forestalled by the community splintering (Molinero, 2000). One or more subgroups simply relocate and set up independent albeit smaller communities. Splintering forestalls the descent into chaos by reducing the population of the group to the point at which it once again is possible to leave the supply of order to the group as a whole. Each

subgroup forgoes the benefits of a large population and relatively more complex economic environment. Smaller numbers and simpler economies, however, insure it against internecine chaos. Splintering, as a conflict resolution technique, seems possible as long as resources are relatively free and abundant, property rights are nonexistent, and emigration costs are low (Molinero, 2000). Splintering allows social order to return as commitments once again gain credibility. Social order results in population growth. Continued population growth allows conflicts once again to arise. This pattern of spin-off and population growth continues. At some point, however, some or all of these communities conclude that they are either unwilling or unable to splinter any further. Our analysis, which began its focus on a community of independent individuals, is now focused on a neighborhood of autonomous communities—none of which, in the presence of the others, has an incentive to limit its population growth. The search for an alternative source of credible commitment begins.

## Solution: Statehood

The state can provide a solution to the difficulties of organizing order. The fundamental exchange underlying the state involves a third party and constituents. In exchange for an exclusive share of the output that attends order, constituents forsake violence as an income-maximizing strategy and centralize violence into the hands of the third party (which can be either an individual or a group of individuals). The state, consequently, is defined as an organization with a comparative advantage in violence extending over a geographical area (North, 1979). Why would the third party take on this task? It is because of the residual—which is the difference between the output that order makes possible to be produced and the sum of output promised to constituents (Alchian & Demsetz, 1972). This difference, guaranteed to the third party by constituents, is maximized when the third party allocates the efficient level of effort toward monitoring and enforcing property rights. This is a problem that the community found difficult to solve previously. Now, when constituents turn to each other and commit to respecting each other's property rights, the third party's presence lends them credibility. Indeed, if anyone were to encroach on the rights of another, the third party is willing and able to punish this behavior. No longer feeling the need to maintain or invest in their violence potential so as to deter or otherwise predate on others, scarce effort is freed up and can be allocated toward enhancing their productivity. One of the fundamental advantages of having a third party structure and enforcing property rights is that economies of scale attend this particular form of specialization or monopolization of this task. It is simply less costly for the state to provide these services rather than the overlapping and redundant actions of community members. In other words, the same degree of exclusivity may be realized at a lower cost or a higher degree of exclusivity at the same cost. By endowing the third party with the motive (i.e., ownership of the residual) and the means (i.e., monopoly on violence) to monitor and enforce property rights, each constituent's dominant strategy becomes cooperation; thus, the gains from social order can be captured.

## Why the State Is Not Sufficient for Economic Prosperity

The value of the state lies in the fact that its representatives have an incentive to supply, monitor, and enforce a set of property rights. Indeed, the state is a mechanism by which citizens can credibly commit to recognizing and respecting the rights of others. However, when contemplating the value of citizenship, an individual considers not only the way in which property rights are structured, which stipulates his or her exclusive right to the gains that attend order, but also their stability. One way to express the idea of stability is to consider it from the perspective of a prospective constituent. Stability is realized if, upon entering the state, no other constituent has an incentive to allocate scarce effort toward violently restructuring your rights in the hopes of realizing a more favorable outcome. Enforcement, of course, is the key concern here. Is the third party willing and able to exclude other constituents and rivals to his or her rule? Because of the residual, we know the third party is definitely willing. Moreover, because of his or her monopoly on violence, the third party is also able to do so. Stable property rights, therefore, are intimately related to the third party's comparative advantage in violence. The lower the opportunity cost he or she confronts when projecting violence, the less costly is it for him or her to exclude others, and the more secure are property rights. However, there is a trade-off. By centralizing violence and in essence volunteering to be coerced, constituents transfer from themselves onto the third party the incentive to employ violence as an opportunistic income-maximizing strategy. Having given up their arms and made the costly investments to enhance their productivity, constituents have become lucrative targets for predation. The third party has the motive (income maximization) and the means (comparative advantage in violence) to act opportunistically and restructure property rights in his or her favor. Indeed, our history is resplendent with countless occurrences of the state using its monopoly on violence to override the very rights it is organized to protect. In other words, while constituents are able to credibly commit to recognizing and respecting each other's property rights, the third party is incapable of such a credible commitment. This is a source of instability whose consequences include prospective constituents reconsidering statehood. Indeed, rational individuals would never agree to a contract that made it incumbent upon them to give up their right to defend themselves when it is possible that once they enter into the contract, another may prey upon them. Necessarily, stability is a function of

the third party's willingness and ability to not only enforce but also credibly commit to property rights.

### Making the State a Sufficient Condition for Economic Prosperity

Barzel (2002), in his theory of the state, addresses the third party's inability to credibly commit and offers the following solution. Prospective constituents, before inviting another individual or group of individuals to take on the role of the third party, construct a "collective action mechanism" with which to protect themselves against the consequent differential in power. Barzel, however, does not provide an explicit example of such a collective action mechanism. North and Weingast (1989) arrived at this conclusion a decade earlier. In their case study of seventeenth-century England, they concluded that power-sharing rules that give rise to political "veto players" can check the king's incentive to arbitrarily restructure property rights. Barzel acknowledges that mechanisms along the lines of North and Weingast are at times incomplete guarantees against the abuse of power; however, he does not address the fact that if prospective constituents are capable of looking ahead and recognizing that the third party is incapable of credibly committing to property rights, then why are they not just as capable of looking ahead and addressing the limitations of their chosen collective action mechanism?

Humphrey (2004) extends Barzel's (2002) analysis by providing an example of such a collective action mechanism. Constituents recognize that by dissipating the third party's comparative advantage in violence, they can diminish the returns that attend predation and thereby blunt the third party's incentive to prey upon them. They can accomplish this by choosing to organize state-sponsored violence in a way that purposively increases the cost of motivating and coordinating the community's scarce resources toward violent ends. For example, they could choose a more decentralized organization of violence that increases the number and changes the character and composition of those who have access to violence (e.g., a citizens militia). Such an organization will increase the third party's costs of employing violence as an opportunistic strategy. Moreover, if the costs are greater than the returns, then it is worthwhile for the third party to commit. Rationality dictates that he or she does. Indeed, his or her commitment is credible. Consequently, stability is also negatively related to the third party's comparative advantage in violence.

The implementation of such a collective action mechanism, however, is attended by a significant trade-off. Indeed, the problem is more complex than originally framed. Earlier we discussed how our original community could use splintering as a conflict mitigation mechanism. Over time, splintering will result in a neighborhood of communities. These communities also confront the same world of scarcity and competition. Realizing a cooperative solution in this context, however, is much more difficult. The conflict mitigating mechanisms of small groups are less effective between groups—for the same reasons enumerated above within a community. Unorganized violence is the likely state of affairs among independent communities. Therefore, when it comes to enumerating those individuals whose actions can influence the stability of a prospective constituent's property rights, we must add to that list external rivals to the third party's rule. Consequently, actively dissipating the third party's comparative advantage in violence can signal weakness to its rivals—enticing them to predate. Although willing to enforce property rights, the third party may no longer be able.

Stability will only be realized when the third party's rivals (both internal and external) recognize and respect his or her constituents' property rights. Prospective constituents must reconsider allowing the third party to maintain a monopoly on the community's violence potential. However, as we already know, this may induce the third party to renege. The risky step centralizing violence will not be acted upon until another collective action mechanism, one that can be substituted for the decentralized organization of violence, can be implemented. For example, instead of the third party unilaterally deciding when, where, and upon whom to project violence, one solution to this tension would be to introduce politics into these decisions. This can be accomplished by increasing the number of individuals who have a say in the use of violence and manipulate their character (popularly elected or appointed) and composition (bicameral or unicameral system). If these constitutional innovations are adopted, then the community can feel comfortable proceeding with the centralization of violence. Moreover, by simultaneously choosing to centralize violence and decentralize the decision to employ violence, the third party may now be capable of signaling strength to his or her rivals and weakness to his or her constituents. Consequently, managing the credible commitment dilemma requires paying attention not only to the number, character, and composition of those who have recourse to violence but also to the number, character, and composition of those who have a say in its use.

## Applications and Empirical Evidence

England's Glorious Revolution of 1688 furnishes an opportunity to evaluate the claims of the aforementioned discussion. North and Weingast (1989) have already demonstrated the role of the Glorious Revolution in establishing the conditions under which the state can credibly commit to property rights. However, there is more to the story—namely, Article VI of the Bill of Rights. Per Article VI of the Bill of Rights, William III, in exchange for the crown, surrendered to Parliament one of the king of England's historical prerogatives: absolute decision-making authority over England's military establishment. The Glorious Revolution, by giving Parliament sovereignty

over state-sponsored violence, created conditions under which the state could raise and maintain a standing army without threatening property rights. Thus, the emergence of Parliamentarians, whose bundle of political rights included a "veto" over military-related decisions, played a pivotal role in enhancing security by allowing the state to place instruments of violence continuously, even in times of peace, in the hands of professionals, thus allowing the state to realize the efficiency-enhancing benefits of a flexible, immediate, and overwhelming response to property right–destabilizing challenges from rivals inside and outside the state—without threatening its own constituents. Pamphlets, authored by political activists during this period of history, provide evidence that constituents were cognizant of the tension surrounding the choice of how to organize violence, of the inadequacies of collective choice mechanisms alone as credible commitment devices, and how their attempts to address this tension had a measurable influence on the political-military structure of their community (Humphrey & Hansen, 2010).

## Future Directions

In economics, we have a definition of the state. The state is an organization with a comparative advantage in violence that extends over a geographic area. For the most part, economists are in agreement regarding the valuable role that the state plays realizing social order. Economists, however, have not adequately explained the process by which the state emerges. That is, we do not fully understand how an individual or group of individuals centralizes and thereby creates a comparative advantage in violence. In theories of the state, the third party is either simply assumed or the story usually told is that self-interested individuals (some subset or the entire population of a community) recognize the aforementioned benefits of centralizing violence in some entity and invite either one of their own or an outsider to specialize in this role. Indeed, looking forward, these individuals recognize that if the conditions are right (the third party has a sustained monopoly on violence), then favorable forces are put into play, which provide the third party with an incentive to not only regularize his or her rate of taxation but also supply ancillary public goods that further facilitate economic growth (North, 1979; Olson, 1993). A number of scholars, however, argue against this process of state formation. Economic agents, they argue, are inherently jealous of their autonomy. The idea that self-interested individuals would willingly subject themselves to a third party is fanciful. In addition, as we have already discussed, prospective citizens will forgo membership until the state can credibly commit to the rules it created. Yet, with respect to social organization, the long-term historical trend has been an increase in the centralization of violence potential. Olson (2000) was correct when he informed us of who the bandits (the "violent entrepreneurs") were—they were the ones "who can organize the greatest capacity for violence" (p. 568). We need to open up the hood of this comment. Is it a function of their ability to finance the minimum efficient scale of military operations? We are not sure. Does one's capacity for violence stem from their physical attributes? In this case, we know the answer is no. It is a biological fact that the weakest can kill the strongest and/or a subset of the weakest can collude and kill the strongest (Molinero, 2000). The process of organizing organized violence—and, in consequence, accumulating and centralizing political authority—still needs to be explored. Here are some things to consider:

1. Collusion among a subset of the weakest requires some degree of cooperation; however, given that we are talking about small groups, the costs of acting collectively—in particular the costs of monitoring each others' behavior—are minimized. Understanding the conditions that allow for and frustrate the emergence of cooperation in the collective act of projecting violence may provide insight into the conditions that allow for and frustrate the emergence of the third party.

2. Earlier we discussed how unorganized violence creates a disincentive to make productivity-enhancing investments. However, it is important to note that those who become more productive may now be able to purchase the violence potential of others. In other words, the relatively more prosperous can also be deadly by co-opting predators into becoming protectors. A constraint on this possibility is that there must exist a low transaction cost environment in the market for violence. Those who are now in the position to purchase the violence potential of others must be able to answer yes when they ask, "Will I get what I paid for?" Conversely, those who are selling their violence potential must be able to answer yes when they ask, "Will I get paid?" Issues of credibility once again arise. The market for violence needs to be explored.

3. Among competing states, war can be a selection mechanism—selecting for survival those states that structure and enforce a system of political and economic property rights that capture the gains from cooperation. Changes in the complexity of the instruments and/or tactics of warfare will require concomitant changes in the organization of state-sponsored violence. For example, in the later part of the seventeenth century, England's citizens militias were coming under increasing criticism for their numerous weaknesses—including disorganization, poor leadership, and outdated tactics and weaponry. All of these defects could in some way be traced back to the militia's extreme decentralization. There were 52 autonomous county militias. These critiques were being leveled while continental powers (in particular, France) were centralizing military power by organizing standing armies in response to changes in military technology and tactics (e.g., the rise

of firearms and the consequent changes in military formations to take advantage of this technology). Political pamphleteers in England argued that survival would require an organizational innovation—that is, a standing army. Adaptations such as these, however, will have an influence on the state's comparative advantage of violence and hence its ability to credibly commit. Concomitant changes may have to take place in the political arena if credibility is to be restored. Necessarily, there may be a link between changes in military technology and constitutional innovations.

In general, understanding the process by which organized violence becomes organized and the factors that influence this process gives us a glimpse into the font of state power and also a glimpse into the ways in which we can tweak said power toward creating credibility.

## Conclusion

Unorganized violence destabilizes property rights, strips individuals of their current wealth and property, and creates an incentive for community members to redirect their efforts from productive growth-augmenting endeavors—for example, specialization and trade—into predatory and/or protective activities. Necessarily, the task set before a community is finding a way to self-regulate the use of violence within its boundaries. If there is an opportunity for mutually beneficial interactions, then individuals will be motivated to take the costly steps to make it happen. In other words, there will be a demand for some rules to govern the use of violence. It is the supply, monitoring, and enforcement of rules that are the fundamental hurdles to the emergence of these rules. A key variable is the willingness and ability of each community member to commit to be bound by the rules. The supply of rules, even under the best conditions (one agent is willing and able to finance the cost of supplying the rules), will be in vain if their commitments lack credibility. The likelihood of accomplishing this task is pretty good when the community is small, homogeneous, and characterized by repeated interactions.

These traits, however, are the very opposite of those traits that characterize a commercial society and form the bedrock of economic prosperity—namely, large numbers, heterogeneous population, a high degree of specialization, and complex trade across space and time. Necessarily, the new task set before the community is to allow for the emergence of an accountable, honest, and responsible individual or group of individuals into whose hands violence is centralized (i.e., organized). It is the state that is usually tasked with providing protection services. Indeed, the value of the third party lies in its willingness and ability to enforce the rules and thereby allow community members to credibly commit to recognizing and respecting each other's property rights. However, in this arrangement, the inability to credibly commit is simply transferred from constituents to the third party. Prospective constituents are aware of this dilemma and will be resistant to monopolizing violence within the third party until they settle upon and introduce mechanisms by which the third party can credibly commit. Until that time, they will forgo the benefits that attend economic prosperity.

The third party's ability to renege and predate is fundamentally a function of his or her violence potential relative to that of his or her constituents. A credible commitment can be engineered by manipulating this power differential through choosing an organization of violence that increases the third party's costs of coordinating and motivating the resources that are set aside for violence (labor, tools, and entrepreneurial spirit). Yet, herein lies a tension. Choosing an organization of violence that is too decentralized, while yielding a credible commitment, makes the third party too weak and may invite insurrection and/or invasion. Choosing an organization of violence that is too centralized, while efficient, makes the third party a threat. A solution to this dilemma requires a constitutional innovation (e.g., Article VI) designed to simultaneously provide for defense while protecting property rights. For example, settle upon a relatively more centralized organization of violence (signal strength to rivals within and outside)—but invest the decision to deploy violence into the hands of political veto players (signal impotence to constituents). Both require manipulating the number, character, and composition of those who have access to violence and those who have a voice in the political arena. Settling upon a particular number, character, and composition in each arena is shaped not only by the dynamic interaction between the third party and constituents but among the third party, his or her constituents, and rivals inside and outside the state.

Successfully managing the credible commitment dilemma, structuring a self-enforcing politico-military settlement, is pivotal to secure property rights. Therefore, truly ascertaining the conditions that underlie economic prosperity requires us to ask the following questions: How do political economies, in response to their strategic environment, organize *organized* violence, and what are the ramifying consequences of their choices? In particular, how do the number, character, and composition of those who have access to violence change in response to a political economy's strategic environment? And, in turn, how do these changes influence the number, character, and composition of those with decision-making authority over a political economy's organization of violence?

## References and Further Readings

Alchian, A., & Demsetz, H. (1972). Production, information costs and economic organization. *American Economic Review, 62,* 777–795.

Andrzejewski, S. (1954). *Military organization and society.* London: Routledge and Paul.

Axelrod, R. (1985). *The evolution of cooperation.* New York: Basic Books.

Barzel, Y. (2002). *A theory of the state: Economic property rights, legal rights and the scope of the state.* Cambridge, UK: Cambridge University Press.

Bates, R. (2001). *Prosperity and violence: The political economy of development.* New York: W. W. Norton.

Collier, P. (2007). *The bottom billion: Why the poorest countries are failing and what can be done about it.* Oxford, UK: Oxford University Press.

Diamond, J. (1999). *Guns, germs and steel: The fates of human societies.* New York: W. W. Norton.

Hirschleifer, J. (1994). The dark side of the force. *Economic Inquiry, 32,* 1–10.

Hobbes, T. (1968). *Leviathan.* London: Penguin. (Original work published 1651)

Humphrey, S. (2004). Protecting your protection in a violent world: The link between a state's organization of violence and its constitutional design. *Homo Oeconomicus, 21,* 117–152.

Humphrey, S., & Hansen, B. (2010). Constraining the state's ability to employ force: The standing army debates, 1697–99. *Journal of Institutional Economics, 6*(2), 243–259.

Miller, G. (1992). *Managerial dilemmas: The political economy of hierarchy.* Cambridge, UK: Cambridge University Press.

Molinero, J. M. S. (2000). The origins of the state from reciprocity to coercive power. *Constitutional Political Economy, 11,* 231–253.

North, D. (1990). *Institutions, institutional change, and economic performance.* Cambridge, UK: Cambridge University Press.

North, D. C. (1979). A framework for analyzing the state in economic history. *Explorations in Economic History, 16,* 249–259.

North, D. C. (1981). *Structure and change in economic history.* New York: W. W. Norton.

North, D. C., & Weingast, B. (1989). Constitutions and commitments: The evolution of institutions governing public choice in 17th century England. *Journal of Economic History, 49,* 803–832.

Nye, J. V. C. (1997). Thinking about the state: Property rights, trade, and changing contractual arrangements in a world with coercion. In J. N. Drobak & J. V. C. Nye (Eds.), *The frontiers of the new institutional economics* (pp. 121–142). New York: Academic Press.

Olson, M. (1965). *Logic of collective action: Public goods and the theory of groups.* Cambridge, MA: Harvard University Press.

Olson, M. (1993). Dictatorship, democracy, and development. *American Political Science Review, 87,* 119–138.

Olson, M. (2000). *Power and prosperity: Outgrowing communist and capitalist dictatorships.* New York: Basic Books.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action.* Cambridge, UK: Cambridge University Press.

Schofield, N. (2000). Institutional innovation, contingency, and war: A review. *Social Choice and Welfare, 17,* 1–17.

Turney-High, H. H. (1971). *Primitive war: Its practice and concepts.* Columbia: University of South Carolina Press. (Original work published 1949)

Umbeck, J. (1981). Might makes rights: A theory of the formation and initial distribution of property rights. *Economic Inquiry, 19,* 38–59.

# 79

# The Economics of Civil War

Marta Reynal-Querol
*Pompeu Fabra University*

Different social sciences have adopted different perspectives in the study of conflict. Most economists have approached this issue using a theoretical perspective and the tools of game theory. During the past two decades, some economists have constructed theoretical models to explain why individuals and groups become involved in a conflictive situation. For an outstanding review on the theoretical literature of the economics of conflict, I strongly recommend reading Garfinkel and Skaperdas (2007).

However, the empirical approach to this topic is still in its early stages. Recently, some researchers have stressed the importance of economic factors in the study of civil war. This chapter will provide a brief overview of the most relevant empirical papers that address this topic from an empirical point of view. This is not an overview of the main empirical literature on civil war but a summary of the most important papers, as well as the evolution and state of the art on this topic, for undergraduate readers.

## Introduction to the Empirical Analysis of Civil Wars

Civil wars have only recently been recognized as one of the main impediments for economic development. Their effects are not only related to the destruction of infrastructure or human life but also to the elimination of the rule of law, the generation of an uncertain environment for future foreign investment, and the destruction of institutions. Empirical research has used different approaches to deal with the explanation of the basic elements of civil wars. Researchers have analyzed the onset of civil wars, the incidence of civil wars, and the duration of civil wars. These indicators are related but measure different concepts. The probability of onset is the conditional probability of being in state $A$ (war) at time $t$ given that it was in state $B$ (peace) at time $t - 1$.

The incidence is the unconditional probability that we observe state $A$ at time $t$. The hazard function measures the probability of being in state $A$ between period $t$ and $t + 1$ given that it was in state $A$ in period $t$. The duration measures the units of time in state $A$ for each event. Obviously, these three analyses are complementary but deal with different sides of the civil war phenomenon.

We can create an analogy with the analysis of macroeconomic cycles. Researchers in that field distinguish between shocks and their propagation mechanism as two different and independently interesting issues. For instance, a cycle could be caused by a productivity shock that is propagated through many alternative mechanisms. In the case of civil wars, the situation is similar, although identification is more difficult.

For instance, in many situations, civil wars start by random acts, which trigger, given a particular propagation mechanism, full-fledged conflicts.

A crucial issue in this topic is understanding the process of conflict.

In order for a conflict to exist, not only do we need an origin that brings about conflict but also a propagation mechanism that allows conflict to exist and propagate. Conflict onset is a highly unpredictable event. In many conflicts, the onset is related to some unpredictable shocks, such as the original trigger of the genocide in Rwanda or the events of 2005 in France (the burning of vehicles during November 2005). All countries receive unexpected shocks, but not all of them enter into conflict.

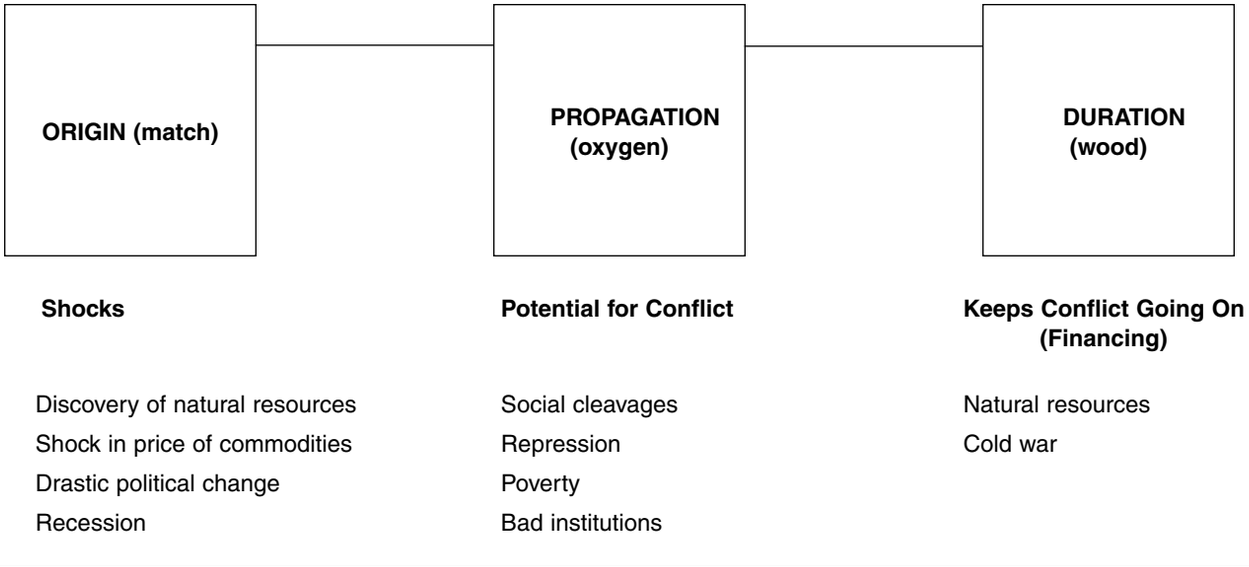


**Figure 79.1** Conflict (Fire)

The countries that do not have potential for conflict—that is, do not have the propagation mechanism for conflict—are the ones that are safe. Therefore, given the process of conflict (shocks, propagation mechanism, and financing), three types of policies could be applied to reduce the probability of conflict: policies that try to avoid shocks (e.g., diversification policies), policies that try to reduce the propagation mechanism (e.g., institutions that try to reduce the intensity of social cleavages), and policies that try to cut the source of financing (e.g., the Kimberly process).

Since onset of conflict is usually produced by unexpected shocks, trying to find measures to prevent them is an impossible task, given the unexpected characteristics of shocks. While diversification can protect countries against shocks concerning the price of some products, it is not possible to predict and avoid unexpected shocks, such as the terrorist attacks on September 11, 2001. Also, policies that address the financing are policies with a very short-term effect because rebel groups look for other alternative sources of financing.

Some of this evidence is explained in the work of Michael Ross (2002), who describes this phenomenon very well with an example of Angola:

> Before the end of the Cold War, successful rebel groups in the developing world were typically financed by one of the great powers. Since the Cold War ended, insurgent groups have been forced to find other ways to bankroll themselves; many have turned to the natural resources sector. . . . In Angola, for example, UNITA (National Union for the Total Independence of Angola) was backed by the United States and South Africa for most of the 1970s and 1980s. But the end of the Cold War, and the end of the apartheid in South Africa, left UNITA with no outside sponsors; as a consequence, it began to rely much more heavily on diamond revenue to support itself. (p. 20)

This is indicative that policies that try to avoid conflict by cutting the sources of financing are policies that have a short-term effect. Therefore, if we want to find measures to prevent conflict in the long run, we need to look for policies that address the propagation mechanism for conflict.

If we want to analyze the cause of civil wars, we need to be aware of what onset, incidence, and duration are capturing, either in panel or cross-sectional data. For example, if we want to analyze the effect of variables that are invariant over time, the most appropriate specification would be to use a cross section of countries.

But before describing the main literature on this topic, we need to clarify what we mean by conflict. Many phenomena can be classified as conflict: riots, demonstrations, coups d'état, political assassinations, terrorist attacks, civil wars, genocide, and mass killing. The empirical literature on the causes of all these conflicts is very recent. This is basically due to the lack of reliable data on conflict. Recently, there has been a huge effort in trying to obtain good data on civil war. We will basically concentrate on the empirical studies that work with civil war data. But then, what is a civil war? A traditional source for data on civil wars is the Armed Conflict Dataset, a joint project between the Department of Peace and Conflict Studies at Uppsala University and the Center for the Study of Civil War at the International Peace Research Institute, Oslo (PRIO). An armed conflict is defined as a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths. This source also provides information on civil war with a threshold of 1,000 battle-related deaths per year.

## Empirical Analysis of the Causes of Civil Wars

The empirical analysis of civil war is very recent and basically concentrates on the analysis of the causes of civil war from a macroeconomic perspective. In particular, the relationship between poverty and civil war from a macroeconomic perspective has been one of the main topics under discussion in the empirical literature on civil war. Now there is a growing emerging literature that addresses these issues from a microeconomic point of view. In this chapter, we will concentrate only on the most developed literature on the causes of civil war from a macro point of view and raise the state of the art on the new development on the microeconomic literature that has just started being developed.

### Relationship Between Poverty and Civil War: The Seminal Work

One of the most robust results in the civil war literature is the correlation between per capita income and civil war. Some studies have interpreted this correlation as causality, but the potential reverse causality and omitted variable problems indicate that we should consider these results with caution. In this section, we will examine the main papers that discuss these issues.

One of the most influential papers in the empirical literature on the relationship between poverty and civil war is Collier and Hoeffler (2004). This essay develops an econometric model that tries to predict the probability of civil war. It focuses on the initiation of conflict, the onset of civil wars. The article argues that any rebellion needs a motive and an opportunity. The political science approach offers an account of conflict in terms of motive: Rebellion occurs because of grievances. The economic approach, inspired by Herschell Grossman (1994) and Jack Hirshleifer (1989, 1991), models rebellion as an industry that generates profits from looting. Such rebellions are motivated by greed. The political science and the economic approaches to rebellion have assumed different motivations: greed versus grievances. Collier and Hoeffler test empirically both models. One of the questions that arises here is how to measure empirically opportunity and grievance.

Opportunities to finance rebellion can be divided into different types: first, natural resources (diamonds, etc.), which are captured by the ratio of primary commodity exports to gross domestic product (GDP); second, diasporas, where the size of a country's diasporas is captured by its emigrants living in the United States; and third, subventions from hostile governments. The proxy for the willingness of foreign governments to finance military opposition to the incumbent governments is cold war. During cold war, each super power supports rebellions in countries allied to the opposing power.

Moreover, recruits must be paid. Rebellions may occur when income is unusually low. Therefore, the paper proxies this opportunity with the mean income per capita, the male secondary schooling, and the growth rate of the economy.

Finally, another opportunity for rebellion is when the conflict-specific capital is cheap. The study uses the time since the most recent previous conflict to proxy this. Because this captures the legacy of weapons stock and so forth, another dimension of opportunity is weak government military capability, which can be captured with the idea that some terrain may be favorable to rebels, such as forests and mountains. Therefore, Collier and Hoeffler (2004) use the following variables to capture these terrain characteristics: percentage of terrain that is forest, percentage of terrain that is mountain, and the dispersion of the population.

Another source of rebel military opportunity is social cohesion. A newly formed army may need social cohesion and can therefore constrain recruitment to a single ethnic group. Therefore, a diverse society might then reduce the opportunity for rebellion.

Collier and Hoeffler (2004) consider different dimensions of grievances: ethnic or religious hatred; political repression (traditional measures of democracy [Polity III]); political exclusion, a measure that captures whether the largest ethnic group constitutes between 45% and 90% of the population; and economic inequality, which is captured by the Gini index.

In the empirical analysis, Collier and Hoeffler (2004) try to predict the risk that a civil war will start during a 5-year period, through a logit regression, during 1965 to 1999. In all of the analysis, the dependent variable is a dummy that has value 1 if a civil war started during the period and zero otherwise. Ongoing wars are coded as missing observations.

The independent variables included to capture economic conditions are per capita income or male schooling. First, Collier and Hoeffler (2004) test the two models separately and then test them together just taking the significant variables. One result of this study is that exports of primary commodities are highly significant. Once primary commodities are disaggregated into different types of commodities (e.g., oil vs. non-oil), the findings are that low levels of oil dependence are less risky than other commodities, and high levels of dependence are more risky. The most significant variables in predicting the probability of civil wars are male secondary education, per capita income (initial period), and growth rate (average of the 5 previous years).

However, we should consider these results with caution since we have endogeneity problems due to reverse causality and omitted variable problems. For example, it could be that the poverty path and civil war path may be caused by some historical determinants, which are missing in the regression. If this is the case, the relationship we observe could be simply spurious.

The best way to address these issues would be to run instrumental variable (IV) regressions for economic growth and per capita income and to include country fixed effects. This is done by other studies that we will describe in the next section.

The second most influential article on civil war that has stressed the relationship between poverty and civil War is Fearon and Laitin (2003). The main question this article tries to answer is, "Why do some countries have more civil wars than others?" Fearon and Laitin use their own definition of civil war, which includes anticolonial wars. One important hypothesis is that financing is one determinant of the viability of insurgency. The authors argue, however, that economic variables such as per capita income matter primarily because they proxy for state administrative, military, and police capabilities. The interpretation is the following: Where states are relatively weak and capricious, both fears and opportunities encourage the rise of would-be rulers who supply a rough local justice while arrogating the power to "tax" for themselves and often for a large cause. Following this argument, the authors test the following hypothesis, which I have summarized directly from Fearon and Laitin's article:

> H1: Measures of a country's ethnic or religious diversity should be associated with a higher risk of civil wars.
>
> H2: The effect of ethnic diversity on the probability of civil war should increase at higher levels of per capita income (a proxy for economic modernization).

The idea is that more modernization should imply more discrimination and thus more nationalist contention in culturally divided societies.

> H3: Countries with an ethnic majority and a significant ethnic minority are at greater risk of civil war.

To capture all these hypotheses, the article uses the following variables: the ethnolinguistic fractionalization index (ELF, which is basically the fragmentation index using the Atlas Narodov Mira data set), the share of the population belonging to the largest ethnic group, the number of different languages spoken by groups exceeding 1% of the country's population, and the religious fractionalization index.

> H4: Measures of political democracy and civil liberties should be associated with the lower risks of civil war onset.
>
> H5: Policies that discriminate in favor of a particular language or religion should raise the risk of civil war onset in states with religious or linguistic minorities.
>
> H6: Greater income inequality should be associated with higher risks of civil war onset.

In this article, the idea is that to explain why some countries have experienced civil wars, one needs to understand the conditions that favor insurgency, which are largely independent of cultural differences between groups.

The idea is that these conditions should be understood in the logic of insurgency. The story explained in the article is the following: The fundamental fact about insurgency is that insurgents are weak relative to the government they are fighting. If the government forces knew who the rebels are and how to find them, they would capture them. Because of the weakness of the insurgents, to survive, the rebels must be able to hide from government forces. Therefore, the following hypotheses are tested:

> H8: The presence of
> - rough terrain, poorly served by roads and at a distance from centers of state power, should favor insurgency and civil war.

So should the availability of

- foreign, cross-border sanctuaries and
- a local population that can be induced not to denounce the insurgents to government agents.

Insurgents are to survive and prosper better if the government and military they oppose are relatively weak (badly financed).

> H9: A higher per capita income should be associated with a lower risk of civil war onset because
>
> - it is a proxy for a state's overall financial, administrative, police, and military capabilities, and
> - it will mark more developed countries with terrain that is more "disciplined" by roads and rural society that is more penetrated by central administration.

They argue that there is another reason why a lower per capita income should favor the technology of insurgency, which is that

- recruiting young men to the life of a guerrilla is easier when the economic alternatives are worse.

It is difficult to find measures to distinguish among these three mechanisms associating a low per capita income with civil war onset. In any case, the main argument provided in the article is that the results of the relationship between per capita income and civil war onset are basically due to the channel of weak states. Fearon and Laitin (2003) believe this occurs because the relationship between the percentage of young males and male secondary schooling rates and civil war is not strong. The empirical analysis uses annual data between 1945 and 1999. The dependent variable is the onset variable, which is a dummy variable that has a value of 1 for the years in which civil war started and zero otherwise (peace and ongoing civil wars). The main results of this study indicate that per capita income, population, and mountains are significant variables, while ethnic fractionalization is not insignificant.

As before, the endogeneity problems due to reverse causality and omitted variable problems imply that we need to be very cautious in the interpretation of these results. Moreover, interpreting why per capita income affects civil war should be solved by analyzing the microeconomics of civil war.

From these two studies, the two main ideas that have emerged and translated into the policy arena are that poverty (per capita income) is the main cause of conflict and that ethnic diversity and democracy do not play any role.

However, these studies have important econometric shortcomings that could reverse the results, once addressed:

1. The current consensus, which emerges from those analyses, is that poverty is the single, most important determinant of civil wars. However, this result could be an artifact of simultaneity problems: The incidence of civil wars and poverty may be driven by the same determinants, some of which are probably missing in the typical econometric specifications. In the next section, we will show what happens once we consider endogeneity problems.
2. They do not address the endogeniety problems between institutions and civil wars.
3. This literature does not use the appropriate index to capture the potential for conflict. Theoretical models of conflict suggest that polarization measures could be better proxies than fractionalization measures. In the next section, we will see what happens once we change the measure to capture ethnic diversity.

All of these three concerns have been recently addressed by the literature. We briefly describe below the strategy followed and the main findings.

## Poverty and Civil War: Addressing Endogeneity Problems

More recently, three studies have begun to address the main concerns regarding the relationship between income and civil wars. We will briefly describe the main innovations of two such studies here (Djankov & Reynal-Querol, in press; Miguel, Satyanath, & Sergenti, 2004) and one in the next section (Besley & Persson, 2008) since they also address the role of institutions.

The first study to address the endogeneity problems between economic variables and civil wars is Miguel et al. (2004). I briefly describe the main innovations and the strategy followed by this study. Therefore, this part is basically based on extracts from Miguel et al.

This article is concerned with civil wars in sub-Saharan African countries. The main motivation for the authors is that the literature highlights the association between economic conditions and civil conflict. However, this literature does not adequately address the reverse causality problem of economic variables to civil war and therefore does not convincingly establish a causal relationship. Also, omitted variables (e.g., institutions) may be driven by both economic outcomes and conflict.

Miguel et al. (2004) use exogenous variation in rainfall as an instrumental variable for economic growth to estimate the impact of economic growth on civil conflict.

They very reasonably argue that weather shocks are plausible instruments for growth in domestic product in economies that largely rely on rain-fed agriculture, do not have extensive irrigation systems, and are not heavily industrialized.

The nature of the econometric identification strategy allows them to focus on short-term economic fluctuations that "trigger" conflicts, but it is not as well suited for understanding conflict duration. But why sub-Saharan African countries? As they argue, "Only one percent of the cropland is irrigated in the median African country, and the agricultural sector remains large. The paper finds that weather shocks are in fact closely related to income growth in Sub-Saharan African countries. This identification strategy is not appropriate for other regions of the world, because weather is not sufficiently closely linked to income growth. The IV approach allows addressing another problem: The measurement error in African national income figures, which are thought to be unreliable" (p. 726).

The econometric specification is basically focused on the incidence of civil war. The dependent variable, incidence, is constructed in the following way: All country-year observations with a civil conflict that has at least 25 battle deaths per year are coded as 1; otherwise, they are coded as zero. The analysis also considers the onset definition (coded 1 for the year in which a civil war starts) since the impact of income shocks on conflict may theoretically differ depending on whether the country is already experiencing conflict. Data on civil war come from Uppsala/PRIO as described above.

The main innovation of this article is to analyze the relationship between economic growth and civil war using an instrumental variable approach. Rainfall is used as an instrumental measure of economic growth. The authors use the Global Precipitation Climatology Project (GPCP) database of monthly rainfall estimates as a source of exogenous weather variation.

There is information on rainfall estimates for each point at which latitude and longitude degree lines cross. This data set measures a latitude-longitude degree node point $p$ in country $i$ during month $m$ of year $t$ : $R_{ipmt}$.

The average rainfall across all points $p$ and months $m$ for that year is $R_{it}$. The principal measure of a rainfall shock is the proportional change in rainfall from the previous year,

$$(R_{it} - R_{it-1})/R_{it-1},$$

denoted as $\Delta R_{it}$.

Weather variation, as captured in current and lagged rainfall growth ($\Delta R_{it}$ and $\Delta R_{it-1}$), is used to measure per capita economic growth $(growth_{it})$ in the first stage, with other country characteristics ($X_{it}$) controlled for.

Country fixed effects ($a_i$) are included in some regressions to capture time-invariant country characteristics that may be related to violent conflict (sometimes also country-specific time trends to capture additional variation).

The main specification of Miguel et al. (2004) is the following:

$$growth_{it} = a_{1i} + X_{it}b_1 + c_{1,0}\Delta R_{it} + c_{1,1}\Delta R_{i,t-1} + d_{1i}year_t + e_{1it}.$$

The main results indicate that current and lagged rainfall growth are both significantly related to income growth. The story behind these results is that positive rainfall growth typically leads to better agricultural production since most of sub-Saharan Africa lies within the semiarid tropics and is prone to drought.

Results for the reduced-form equation estimate the effect of rainfall growth on civil conflict. Higher levels of rainfall are associated with significantly less conflict in the reduced-form regression for all civil conflicts. This indicates that better rainfall makes civil conflict less likely in Africa.

A second-stage equation estimates the impact of income growth on the incidence of violence:

$$conflict_{it} = \alpha_{2i} + X_{it}\beta_2 + \gamma_{2,0}growth_{it} + \gamma_{2,1}growth_{it-1} + \delta_{2i}year_t + \varepsilon_{2it}.$$

One way of looking at the endogeneity problem is to run an instrumental variable estimation for civil wars, disregarding the fact that this is a 0–1 variable that is an IV–two-stage least square (2SLS). Angrist (1991) shows that if we ignore the fact that the dependent variable is dichotomous and use the instrumental variables approach, the estimates are very close to the average treatment effect obtained using a bivariate probit model. Moreover, following Angrist and Krueger (2001), the IV-2SLS method is typically preferred even in cases in which the dependent variable is dichotomous.

The main results of the ordinary least squares (OLS) and Probit analyses indicate that that lagged economic growth rates are negatively but not statistically significantly correlated with the incidence of civil war with country controls. However, the IV results indicate that lagged economic growth has a negative and statistically significant effect on the incidence of civil war (using the definition of more than 25 deaths per year). Moreover, contemporaneous economic growth has also a negative and statistically significant effect on the incidence of civil war (more than 1,000 deaths per year).

Let us now provide some discussion on the exclusion restriction, which Miguel et al. (2004) raised in their article:

> While it is intuitive that rainfall instrument is exogenous, it must also satisfy the exclusion restriction: Whether shocks should affect civil conflict only through economic growth. Can we think of a possible channel through which rainfall shocks affect conflict apart from economic growth? It could be that high levels of rainfall might directly affect civil conflict independently of economic conditions. For example, floods may destroy roads and then make it more costly for government troops to contain rebellion. Since in the reduced form we observe that higher levels of rainfall are empirically associated with less conflict, then we should not worry too much about this possibility. (p. 745)

Also, it could be that rainfall may make it more difficult for both government and rebel forces to engage each other in combat and to achieve the threshold number of deaths that constitutes a conflict. To explore this possibility, they estimate the impact of rainfall shocks on the extent of the usable road network and do not find a statistically significant relationship.

In the main conclusions of their article, Miguel et al. (2004) argue that the results support previous findings on the literature, particularly the results of Collier and Hoeffler (2004) and Fearon and Laitin (2003), who argue that economic variables are often more important determinants of civil wars than measures of grievances. "Collier and Hoeffler stress the gap between returns from taking arms relative to those from conventional economic activities, such as farming, as the causal mechanism linking low income to the probability of civil war. Fearon and Laitin argue that individual opportunity cost matters less than state military strength and road coverage. They argue that low national income leads to weaker militaries and worse infrastructure and thus make it difficult for poor governments to repress insurgencies" (Miguel et al., 2004, p. 728). Miguel et al. argue that the results are consistent with both explanations, and they view the opportunity cost and the repressive state capacity as complementary rather than competing explanations.

However, the analysis by Miguel et al. (2004) is on shocks, changes of per capita income on conflict, not on the relationship between poverty (level) and civil war. Djankov and Reynal-Querol (in press) address the relationship between poverty and civil war. They find that their correlation is spurious and is accounted for by historical phenomena that jointly determine income evolution and conflict in the post–World War II era. In particular, the statistical association between poverty, as proxied by income per capita, and civil wars disappears once we include country fixed effects. Also, using cross-sectional data for 1960–2000, they find that once historical variables such as European settler mortality rates, population density in the year 1500, and European settlement in 1900 are included in the civil war regressions, poverty does not have an effect on civil wars. These results are confirmed using longer time series for 1825 to 2000. The results are in line with Krueger and Malecková (2003), who provide evidence that any relationship between poverty and terrorism is indirect. Abadie (2006) also shows that terrorist risk is not significantly higher in poorer countries once the effect of other

country-specific characteristics, such as the level of political freedom, is taken into account.

These results can be consistent with Miguel et al. (2004), who find that sudden changes in income growth affect the probability of conflict. They analyze the effect of one component of income growth, transitory shocks caused by the change in rainfall. One can imagine a situation where a sudden (and exogenous) hit in consumption drives people to violence. Once various such effects cumulate to increase or reduce the level of income, the effect on civil war seems to disappear. The relationship is similar to the relationship between income and democracy, as well as economic growth and democracy. Acemoglu and Robinson (2001), for example, emphasize that regime changes are more likely during recessionary periods because the costs of political turmoil, both to the rich and to the poor, are lower during such episodes. This is analogous to the results of Miguel et al. However, Acemoglu and Robinson also find that "holding inequality and other parameters constant, rich countries are not more likely to be democratic than poor ones" (p. 949). This is analogous to the result in Djankov and Reynal-Querol (in press).

In Djankov and Reynal-Querol's (in press) recent study, the explanatory variables follow the basic specifications of the literature on civil war, particularly Collier and Hoeffler (2004) and Fearon and Laitin (2003). Collier and Hoeffler consider population size an additional proxy for the benefits of a rebellion since it measures potential labor income taxation. Fearon and Laitin indicate that a large population implies difficulties in controlling what goes on at the local level and increases the number of potential rebels that can be recruited by the insurgents.

The basic specification used in Djankov and Reynal-Querol (in press) is therefore

$$cw_{it} = \alpha ly_{i(t-1)} + \beta lpop_{i(t-1)} + X'_{i(t-1)}\gamma + \delta_t + \lambda_i + \varepsilon_{it},$$

where $cw_{ti}$ is a dummy that has a value of 1 if there is a civil war in the country and zero otherwise, $ly_{i(t-1)}$ is the lagged value of the natural log of per capita income, $lpop_{i(t-1)}$ is the lagged value of the log of population, $X$ is a vector of all other potential covariates, and $\delta_t$ denotes the full set of time effects that capture common shocks or trends to the civil wars of all countries. They include a full set of country dummies in $\lambda_i$. Finally, $\varepsilon_{it}$ is an error term.

The standard regression in the literature (see previous section) usually omits country fixed effects ($\lambda_i$). In this context, these dummies capture any time-invariant country characteristics that affect the probability of civil war. This is important in the study of the relationship between per capita income and civil war, as some determinants that affect the condition for conflict may at the same time be the condition for economic development.

Djankov and Reynal-Querol (in press) first replicate the results reported in the previous literature. They perform a pooled OLS estimation of the effect of per capita income on the incidence and onset of civil war, using panel data from 1960 to 2000. They use three definitions of civil war. First, they use the definition of the incidence of civil war, which corresponds to more than 25 battle-related deaths per year; the definition of the onset of civil war from the Armed Conflict Dataset, which corresponds to more than 1,000 battle-related deaths in at least one year; and the 1,000-death threshold for the definition of the incidence of civil wars. All regressions include time dummies, and all have robust standard errors clustered at the country level. The results are in line with the literature and show that per capita income has a negative and significant effect on the probability of civil war, whether they use the incidence variable or the onset variable. Also, the results are robust to the use of different thresholds for the definition of civil wars. Then they perform the same analysis but control for time-invariant country-specific variables. Results show that the relationship between per capita income and civil war disappears once fixed country effects are included.

The results using fixed country effects indicate that the relationship between income and civil war is possibly spurious. It is likely that the colonization strategies brought by Europeans were important determinants for the economic development and political stability paths taken by colonies. Using a cross section of countries from 1960 to 2000, they show that while the effect of per capita income on civil war is robust to the inclusion of some contemporaneous variables, its effect disappears once they include historical variables that capture colonization strategies. In the cross-sectional specification, the dependent variable is a dummy that has a value of 1 if the country suffered a civil war during 1960 to 2000 and zero otherwise. In order to reduce the endogeneity problems between per capita income and civil war, the independent variables are taken at the beginning of the period. The specification is

$$cw_{i60-00} = \alpha + \beta_1 \lg dp_{i60} + \beta_2 lpop_{i60} + X'_{i60}\phi + \varepsilon_i,$$

where $cw$ is a dummy variable that has a value of 1 if the country had a civil war during 1960 to 2000 and zero otherwise, $\alpha$ is a constant, lg $dp$ is the log of real per capita income in 1960, $lpop$ is the log of the population of the country in 1960, and $X$ is a set of covariates, some of which are time invariant. All regressions have robust standard errors.

All these results indicate that the correlation between poverty and civil war could be accounted for by historical phenomena that jointly determine income evolution and conflict in the post–World War II era. A plausible explanation for the results found in the literature is that some determinants favor both economic development and peaceful negotiations, which are absent in the traditional specification. If this is the case, then OECD countries are peaceful not because they are rich but because historically they have benefited from circumstances that have favored negotiated settlements and economic development at the same time.

The historical variable could be a proxy of this historical phenomenon that jointly determines the income and conflict path of countries. In another study, explained below, Djankov and Reynal-Querol (2007) propose that institutions, determined by historical variables, could be the channel that determines the path that countries follow.

## Poverty, Institutions, and Civil War

To date, very few and recent studies in economics link economic development, institutions, and civil wars. Djankov and Reynal-Querol (2007) investigate whether the quality of economic institutions has played a role in sustaining peace. In particular, they test the hypothesis that when governments cannot enforce the law and protect property rights, conflict emerges. The idea that strong institutions prevent conflict derives from the theoretical literature of conflict: Haavelmo (1954), Grossman (1994; Grossman & Kim, 1996), Skaperdas (1992, 1996), Garfinkel (1990), and Hirshleifer (1995), among others. The Djankov and Reynal-Querol (2007) study is also related to the extensive empirical literature that has investigated the role of institutions in development that shows a positive relationship between institutions and various proxies for development. Their empirical approach is closely related to this literature. The common idea in the literature is that some historical roots are based on the European influence during colonization that explain institutional development and have nothing to do with contemporaneous factors—in our case, civil wars. They follow the work of Acemoglu et al. (2001), who propose a theory of institutional differences among countries colonized by Europeans, based on the role of settler mortality in shaping local institutions. Consistent with Acemoglu et al., they also study institutional differences between British colonies and colonies from the other major imperial powers (France, Spain, and Portugal).

The results indicate that a lack of secure property rights and law enforcement is a fundamental cause of civil war. Moreover, once institutions are included in the regression analysis, income does not have any direct effect on civil war. This suggests that the direct effect of per capita income found in previous literature may have simply captured the effect of institutions.

Recently, Besley and Persson (2008) have provided a theoretical and empirical framework for the study of the causes of conflict. The aim of this study is to develop a theoretical model on the economic and institutional determinants of civil war and to use this model to interpret the evidence on the prevalence of civil conflict across countries and its incidence within countries over time. This work is a first step along an iterative path where the development of theory and empirical work in this area is joined together. The main empirical contribution of this study is that it looks at the incidence of conflict, controlling for unobserved causes behind the uneven incidence of civil war across countries and time by fixed country effects and fixed year effects. One of the main results of this study is that country-specific price indices constructed for agricultural products, minerals, and oils have considerable explanatory power in predicting the within-country variation of conflict. Preliminary results indicate that higher prices of exported commodities raise the probability of observing conflict. The same result is found for a higher process of imported commodities. This seems to depend on the institutional framework of the country.

Besley and Persson (2008) build a model that serves as a useful guide for how observable economic and political factors determine the probability of violent domestic conflict. This model gives a transparent set of predictions on how parameters of the economy and the polity affect the incidence and severity of conflict.

Besley and Persson (2008) develop a simple microfounded model to illustrate how prices of importable and exportable commodities affect wages and natural resource rents and hence the incidence of civil war over time. The story is as follows: A higher price of the imported raw material lowers the wage, which raises rents in the export sector and hence the prize for winning conflict. The lower wage also has a direct positive effect on the probability of observing conflict by lowering the opportunity cost of fighting and hence conflict. The economic intuition behind these results is that higher prices for exported commodities have a direct effect on civil war by increasing rents. The effect of higher imported commodity prices comes from the fact that they reduce the demand for labor in the importable sector and hence put downward pressure on the wage.

This study also estimates panel regressions with a binary civil war indicator as the dependent variable and with fixed country effects. In this way, the analysis can identify the effect of resource rents and real income on the incidence of civil war exclusively from the within-country variation of these variables. This is in contrast to the existing empirical literature that does not usually include country fixed effects.

Besley and Persson (2008) wish to exploit changes in commodity prices in world markets to generate exogenous time variation in resource rents and real incomes. They construct country-specific export price and import price indices. Given the prediction of the model, they interpret a higher export price index as a positive shock to natural resource rents, as well as a higher import price index as a negative shock to income. The empirical results indicate that both export and import indices for agricultural and mineral products are positively and significantly correlated with the incidence of civil war. They also show that the effects of the world market process are heterogeneous, depending on whether a country is a parliamentary democracy or has a system of strong checks and balances.

## Ethnic Diversity and Civil War

As we explain above, the literature on the study of the relationship between ethnic diversity and conflict uses

ethnic fractionalization measures to capture the effect of ethnic grievances on civil war. The findings are that ethnic diversity has no effect on civil war (using ethnic fractionalization measures).

Several authors have stressed the importance of ethnic heterogeneity in the explanation of growth, investment, and the efficiency of government for civil wars. Easterly and Levine (1997) find empirical evidence to support their claim that the very high level of ethnic diversity of countries in Africa explains a large part of their poor economic performance. Several authors have interpreted the finding of a negative relationship between ethnic diversity and growth as being a consequence of the high probability of conflict associated with a highly fractionalized society. For this reason, many studies use ELF as the indicator of ethnic heterogeneity. The raw data for this index come from the *Atlas Narodov Mira* (Bruck & Apenchenko, 1964) compiled in the former Soviet Union in 1960. The ELF index was originally calculated by Taylor and Hudson (1972). In general, any index of fractionalization can be written as

$$FRAG = 1 - \sum_{i\ 1}^{N} \pi_i^2,$$

where $\pi_i$ is the proportion of people who belong to the ethnic (religious) group *i,* and *N* is the number of groups. The index of ethnic fractionalization has a simple interpretation as the probability that two randomly selected individuals from a given country will not belong to the same ethnic group.

However, many authors have found that, even though ethnic fractionalization seems to be a powerful explanatory variable for economic growth, it is not significant in the explanation of civil wars and other kinds of conflicts. These results have led many authors to disregard ethnicity as a source of conflict and civil wars. Fearon and Laitin (2003) and Collier and Hoeffler (2004) find that neither ethnic fractionalization nor religious fractionalization has any statistically significant effect on the probability of civil wars.

However, it is not clear to what extent an index of diversity could capture potential ethnic conflict. In principle, claiming a positive relationship between an index of fractionalization and conflict implies that the more ethnic groups there are, the higher is the probability of a conflict. Many authors would dispute such an argument. Horowitz (1985), who is the seminal reference on the issue of ethnic groups in conflict, argues that the relationship between ethnic diversity and civil wars is not monotonic: There is less violence in highly homogeneous and highly heterogeneous societies and more conflicts in societies where a large ethnic minority faces an ethnic majority. If this is so, then an index of polarization should better capture the likelihood of conflicts or the intensity of potential conflict than an index of fractionalization.

Montalvo and Reynal-Querol (2005) argue that one possible reason for the lack of explanatory power of ethnic heterogeneity on the probability of armed conflicts and civil wars is the measure for heterogeneity. In empirical applications, researchers should consider a measure of ethnic polarization, the concept used in most of the theoretical arguments, instead of an index of ethnic fractionalization.

$$\text{Polarization} = RQ = 1 - \sum_{i\ 1}^{N} \left( \frac{0.5 - \pi_i}{0.5} \right)^2 \pi_i$$

The original purpose of this index, constructed by Reynal-Querol (2002), was to capture how far the distribution of the ethnic groups is from the (1/2, 0,0, . . . 0,1/2) distribution (bipolar), which represents the highest level of polarization. This type of reasoning is frequently present in the literature on conflict. Montalvo and Reynal-Querol (2005) show how to obtain the RQ index from a pure contest model, particularly on ethnic conflict. Esteban and Ray (1999) show, using a behavioral model and rather a general metric of preferences, that a two-point symmetric distribution of population maximizes conflict.

Montalvo and Reynal-Querol (2005) analyze the empirical support for the link between ethnicity and conflict. They pursue this objective by reexamining the evidence on the causes of civil wars using alternative indices to measure ethnic diversity. The empirical section of this article shows that the index of ethnic polarization is a significant explanatory variable for the incidence of civil wars. This result is robust to the use of other proxies for ethnic heterogeneity, alternative sources of data, and the use of a cross section instead of panel data. Therefore, it seems that the weak explanatory power of ethnic heterogeneity on the incidence of civil wars found by several recent studies is due to the use of an index of fractionalization instead of an index of polarization.

Montalvo and Reynal-Querol (2005) estimate a logit model for the incidence of civil wars as a function of polarization and fractionalization measures of ethnic and religious heterogeneity. The sample includes 138 countries from 1960 to 1999 and is divided into 5-year periods. The endogenous variable is the incidence of a civil war. The data on civil wars come from the PRIO data set, particularly the definition of intermediate and high-intensity civil wars of PRIO.

In the empirical analysis, the explanatory variables for the core specification of the incidence of civil wars include the log of real GDP per capita in the initial year (LGDPC), the log of the population at the beginning of the period (LPOP), primary exports (PRMEXP), mountains (MOUNTAINS), noncontiguous states (NONCONT), and the level of democracy (DEMOCRACY).

Results show that the index of ethnolinguistic fractionalization (ETHFRAC) has no statistically significant effect on the incidence of civil wars, following the results of the previous literature. This result is consistent with Fearon and Laitin (2003) and Collier and Hoeffler (2004). However, once they substitute the index of ethnic fractionalization with the RQ index of ethnic polarization, ETHPOL, they find a positive and statistically significant effect on the incidence of civil wars. These results are robust to

the inclusion of alternative measures of heterogeneity such as ethnic dominance or a large ethnic minority. Besides the indices of fractionalization and polarization, the literature has proposed some other indicators of potential ethnic conflict. Collier (2001) notices that ethnic diversity could not only be an impediment for coordination but also an incitement to victimization. Dominance, or one ethnic group in the majority, can produce victimization and therefore increase the risk of a civil war. Therefore, the effect of ethnic diversity will be conditional on being measured as dominance or fractionalization. In principle, fractionalization should make coordination more difficult and, therefore, civil wars will be less probable since it will be difficult to maintain cohesion among rebels. The empirical results reported by Collier seem to indicate that a good operational definition of dominance implies a group that represents between 45% and 90% of the population. However, Collier and Hoeffler (2004) find that dominance, defined as mentioned above, has only a weak positive effect on the incidence of civil wars. When ethnic dominance is included with the RQ index, its coefficient is not significant, while ethnic polarization continues being a significant explanatory variable on the probability of civil wars. Caselli and Coleman (2006) propose another indicator, which is the product of the largest ethnic group (ETHLRG) by primary exports (PRIMEXP). This variable has a coefficient that is not significantly different from zero. The index of polarization is significant even when the product of the largest ethnic group by primary exports is included as an explanatory variable. Finally, the article could also have included the size of the largest minority (LARGMINOR) as another way to proxy polarization. The coefficient of this new variable is not statistically significant, while ethnic polarization continues to be significant even in the presence of this new variable.

The results of Montalvo and Reynal-Querol (2005) are robust to including dummy variables for the different regions of the world. They are also robust to the elimination of regions that are considered especially conflictive, the use of different operational definitions of civil war, the use of alternative data sources to construct ethnic polarization, and the use of cross-sectional samples. Given the nature of the ethnic measures, which are time invariant, they perform regressions in a cross section. In this case, the dependent variable now takes the value of 1 if a country has suffered a civil war during the whole sample period (1960–1999) and zero otherwise.

## Future Development

The study of the economic causes of civil war is still in its early stages. The main studies have been from a macroeconomic perspective, and more has to be done to understand the mechanisms that explain why, for example, income shocks affect conflict. This line of research is little explored, but Dube and Vargas (2008) have taken an important step in this direction in a recent manuscript titled *Commodity Price Shocks and Civil Conflict: Evidence From Colombia.* They try to answer how income shocks affect armed conflict. This article exploits exogenous price shocks in international commodity markets and a rich data set on civil war in Colombia to assess how different income shocks affect conflict. Also, more research needs to be done on the institutional and social mechanism and its interaction with economic shocks that are related to violence.

## References and Further Readings


Abadie, A. (2006). Poverty, political freedom, and the roots of terrorism. *American Economic Review Papers and Proceedings, 96*(2), 50–56.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review, 91,* 1369–1401.

Acemoglu, D., & Robinson, J. (2001). A theory of political transition. *American Economic Review, 91,* 938–963.

Angrist, J. D. (1991). *Instrumental variables estimation of average treatment effects in econometrics and epidemiology* (NBER 0115). Cambridge, MA: National Bureau of Economic Research.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variable and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives, 15,* 69–85.

Besley, T., & Persson, T. (2008). *The incidence of civil war: Theory and evidence* (NBER Working Paper 14585). Cambridge, MA: National Bureau of Economic Research.

Bruck, S. I., & Apenchenko, V. S. (Eds.). (1964). *Atlas Narodov Mira* [Atlas of the people of the world]. Moscow: Glavnoe Upravlenie Geodezii i Kartografii,

Caselli, F., & Coleman, W. J., II. (2006). *On the theory of ethnic conflict* (NBER Working Paper 12125). Cambridge, MA: National Bureau of Economic Research.

Collier, P. (2001). Implications of ethnic diversity. *Economic Policy, 16,* 127–166.

Collier, P., & Hoeffler, A. (2004). Greed and grievances in civil wars. *Oxford Economic Papers, 56,* 563–595.

Djankov, S., & Reynal-Querol, M. (2007). *The causes of civil war* (World Bank Policy Research Working Paper No. 4254). Washington, DC: World Bank.

Djankov, S., & Reynal-Querol, M. (in press). Poverty and civil war: Revisiting the evidence. *Review of Economics and Statistics.*

Dube, O., & Vargas, J. (2008). *Commodity price shocks and civil conflict: Evidence from Colombia.* Unpublished manuscript, Harvard University, Cambridge, MA.

Easterly, W., & Levine, R. (1997). Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics, 112,* 1203–1250.

Esteban, J., & Ray, D. (1999). Conflict and distribution. *Journal of Economic Theory, 87,* 379–415.

Fearon, J., & Laitin, D. (2003). Ethnicity, insurgency, and civil war. *American Political Science Review, 97,* 75–90.

Garfinkel, M. R. (1990). Arming as a strategic investment in a cooperative equilibrium. *American Economic Review, 80,* 50–68.

Garfinkel, M. R., & Skaperdas, S. (2007). Economics of conflict: An overview. In T. Sandler & K. Hartley (Eds.), *Handbook of defense economics* (Vol. II, pp. 649–709). Amsterdam: North-Holland.

Grossman, H. I. (1994). Production, appropriation, and land reform. *American Economic Review, 84,* 705–712.

Grossman, H. I., & Kim, M. (1996). Predation and accumulation. *Journal of Economic Growth, 1,* 333–351.

Haavelmo, T. (1954). *A study in the theory of economic evolution.* Amsterdam: North-Holland.

Hirshleifer, J. (1989). Conflict and rent-seeking success functions: Ratio vs. difference models of relative success. *Public Choice, 63,* 101–112.

Hirshleifer, J. (1991). The paradox of power. *Economics and Politics, 3,* 177–200.

Hirshleifer, J. (1995). Anarchy and its breakdown. *Journal of Political Economy, 103,* 26–52.

Horowitz, D. (1985). *Ethnic groups in conflict.* Berkeley: University of California Press.

Keen, D. (1998). *The economic functions of violence in civil wars* (Adelphi Paper 320). London: International Institute of Strategic Studies.

Krueger, A., & Malecková, J. (2003). Education, poverty and terrorism: Is there a causal connection? *Journal of Economic Perspectives, 17,* 119–144.

Le Billon, P. (2001). The political ecology of war: Natural resources and armed conflicts. *Political Geography, 29,* 561–584.

Miguel, E., Satyanath, S., & Sergenti, E. (2004). Economic shocks and civil conflicts: An instrumental variables approach. *Journal of Political Economy, 112,* 725–753.

Montalvo, J. G., & Reynal-Querol, M. (2005). Ethnic polarization, potential conflict, and civil wars. *American Economic Review, 95,* 796–816.

Reynal-Querol, M. (2002). Ethnicity, political systems and civil war. *Journal of Conflict Resolution, 46,* 29–54.

Ross, M. (2002). *Natural resources and civil war: An overview with some policy options.* Retrieved from http://siteresources.worldbank.org/INTCPR/1091081-1115626319273/20482496/Ross.pdf

Skaperdas, S. (1992). Cooperation, conflict, and power in the absence of property rights. *American Economic Review, 82,* 720–739.

Skaperdas, S. (1996). Contest success functions. *Economic Theory, 7,* 282–290.

Taylor, C., & Hudson, M. C. (1972). *The world handbook of political and social indicators* (2nd ed.). New Haven, CT: Yale University Press.

# 80

# Economic Aspects of Cultural Heritage

Leah Greden Mathews
*University of North Carolina at Asheville*

Cultural heritage is universal in that every culture has a heritage, but that heritage is unique to each culture or community. A building that may be slated for replacement in one region because it is mundane or out of date may be revered in another due to the cultural meaning attached to the building's heritage. As a result, different preferences for cultural goods arise from the differences in culture. In addition, one rarely hears that we have "too much" cultural heritage; more heritage is universally desired, but protecting and creating cultural heritage is costly. Thus, there are many reasons for the study of the economics aspects of cultural heritage, including a desire to study the values that people have for cultural heritage as well as to inform the efficient management of cultural heritage assets.

The economics of cultural heritage is, like many applied economic fields, the result of the application of microeconomic and macroeconomic concepts to the study of a particular facet of our lives, cultural heritage. Cultural heritage has both tangible and intangible aspects, and thus the economics of cultural heritage invokes methodologies used to study both market and nonmarket goods and services.

This chapter begins with a definition of cultural heritage and discussion of the scope of the economics of cultural heritage and next provides an overview of the primary theoretical issues addressed by the economics of cultural heritage. The chapter then moves to discuss empirical examples and applications and policy implications. The fifth section offers emerging trends in the economics of cultural heritage and identifies gaps in the literature where opportunities for significant contributions to the economics of cultural heritage exist for twenty-first-century researchers. The final section offers a summary.

## What Is the Economics of Cultural Heritage?

### Definition

Cultural heritage includes stories, collections, and other artifacts that are used to define and convey the specific attributes of a culture. Thus, cultural heritage is the set of tangible and intangible assets that help to uniquely define a community or nation. Vaughan (1984) indicated that a nation's cultural heritage included three distinct types: the artistic, the natural, and the historical. Some heritage assets are constructed; these include architecture, archaeological sites, and monuments, which are tangible assets, as well as cultural goods such as art, songs, dance, and stories that may be intangible or ephemeral. In addition, some heritage assets are natural assets such as trees imbued with cultural meaning such as the California redwoods; these assets would exist without human intervention, but the values assigned to them are humanly constructed and help to define a culture. Thus, we may think of a country's or region's cultural heritage as a type of capital asset that includes both natural (trees or landscapes) and built assets (monuments or archaeological sites), some of which are tangible (buildings) and some of which are intangible (customs).

A significant challenge exists for those studying the economics of cultural heritage since the definition of cultural

heritage not only is broad but also likely includes distinctly subjective elements. As a result of the diversity of heritage assets and the fact that they are, by definition, uniquely defined for each region, the methodologies used to study the economics of cultural heritage are also diverse.

### The Scope of the Economics of Cultural Heritage

The economics of cultural heritage is often situated as a theme in the cultural economics subdiscipline, which includes the study of cultural industries such as the art and music markets and the consumption and production of cultural goods such as film, art, music, and books. Both cultural economics and the economics of cultural heritage investigate the role of government in the provision of cultural assets, including the study of subsidy, tax, and other policies in the provision or protection of these assets. The economics of cultural heritage also has many overlaps with the environmental and natural resource economics literature due to the nature of the assets being studied (unique, place based), the methodologies used to study them (nonmarket valuation), and the shared characteristics of public goods and externalities that motivate a role for government intervention in both cases.

Outside of economics, the economics of cultural heritage has several important links to sociology and anthropology, especially the sociology of art, which studies how cultures create value. Due to the fact that many cultural heritage sites are historic in nature, the economics of cultural heritage rubs elbows with the fields of history and historic preservation, architecture, and urban planning. And because many heritage sites are tourist attractions, the economics of cultural heritage also often finds itself aligned with tourism studies.

## Theory

Throsby (1997) provides an excellent overview of the core concerns in the economics of cultural heritage. At its most basic level, the first step for any study in the economics of cultural heritage is to clearly define the good that is being studied. In the case of cultural heritage, this is less straightforward than in many other applied economic fields. Once defined, the central theoretical concern in the economics of cultural heritage is clearly the question of the value of cultural heritage. Because of the public goods nature of cultural heritage assets, the role of government intervention is an important theme as well. To study these elements, researchers in the economics of cultural heritage apply several different theoretical constructs from economics.

### The Nature of Cultural Heritage

#### *Public Goods*

Many cultural heritage assets have been identified as having the public good characteristics of nonexclusion and shared consumption. Nonexclusion implies that it is difficult if not impossible to exclude nonpayers from enjoying the benefits of the good. In the case of many monuments and buildings with architectural elements relevant to cultural heritage, it is clear that nonexclusion applies. Shared consumption means that multiple consumers can enjoy the same good simultaneously without reducing the benefit to any one individual. This, too, is relevant for some cultural heritage assets as many consumers can enjoy viewing the exterior façade of the Cathedral of Notre Dame or an archaeological site at the same time.

The economics of market failure tells us that due to the rational free-riding behavior of consumers when public goods are present, these goods will not be provided in socially optimal levels by markets. It is this market underprovision that invokes a potential role for the government provision of public goods. The appropriate role of government in the provision and protection of cultural heritage is thus a key topic for the economics of cultural heritage. Questions regarding whether the destruction of cultural heritage assets should be regulated are studied by cultural heritage economists, as is the determination of the socially optimal amount of investment in a region or nation's cultural heritage assets. Once government has invested in cultural heritage assets as theory predicts, another question for the economics of cultural heritage is whether that investment actually provides social benefits that outweigh the costs of the protection.

#### *Externalities*

Externalities are said to exist when economic actors other than those directly involved in the market transaction are affected by the production or consumption of the good. Second-hand smoke is a classic example of a negative externality since innocent bystanders are negatively affected from the actions of the smoker. When negative externalities are present, market outcomes are not efficient since the market provides more than the socially optimal amount of the good. Positive externalities also exist when bystanders receive benefits from the actions of others. When positive externalities are present, the market will provide less than the socially optimal amount of good.

There are many cases where positive externalities exist with cultural heritage assets. One example can be found in the protection of architectural elements in historic districts that lead to the increased prices of neighboring properties. Similarly, if an individual has protected an old mill or barn from destruction by maintaining the property on his or her own land, the entire community will benefit from that protection since the cultural heritage asset is being maintained as a visual reminder of the historic, symbolic, and perhaps aesthetic values of regional culture.

Many cultural heritage assets are provided by private parties, of course. Private parties regularly maintain cultural heritage sites such as cemeteries without the assistance of government intervention. And in addition to fine

examples of architectural heritage that may be owned by private individuals, Native American rock drawings and burial grounds may be found on what is now private property, due to the history of land transfers. The protection of those assets may be optimal for the collective good, but private individuals will tend to under invest in their protection given the nature of externalities. As a result, the inventory of cultural heritage assets may decline over time.

Government intervention in markets where externalities are present comes in many forms. One form of government intervention is the regulation of activities that promote positive externalities or discourage negative externalities. In the instance of cultural heritage, governments regulate historic districts and prohibit the destruction, sale, or commercial use of certain kinds of cultural artifacts. However, it is difficult to regulate cultural heritage if governments are not aware of the assets, as is sometimes the case when the heritage assets are completely contained on private property.

Another form of government intervention in markets with externalities is to use fiscal policy to encourage (or discourage in the case of negative externalities) the activities. Subsidies for some historic preservation work exist in the form of government grants, and tax credits are granted for renovations and protection work in federally designated historic districts.

## The Value of Cultural Heritage

By far and away the most central concern of the economics of cultural heritage has been the study of the value of cultural heritage. The treatment of value in economics has an interesting evolution, as succinctly described by Throsby (2001, pp. 20–23). A good's market price is typically considered the most effective indicator of economic value that we can identify. The so-called paradox of value, which questions why diamonds, a nonessential luxury, are more expensive than water, a necessity, points to an important caution for relying on market prices as indicators of value. It is clear that for many goods, values other than economic value are not likely to be reflected in market prices; these include cultural values. As a result, a student of the economics of cultural heritage must recognize that market prices cannot directly measure the total value of a cultural heritage asset.

A holistic measure of the value of a cultural heritage asset would include all elements of its cultural value—that is, its historic, aesthetic, spiritual, social, symbolic, and authenticity values (Throsby, 2001) in addition to its economic value. The *historic values* associated with a cultural heritage asset may be readily conveyed by its mere existence, which provides a direct connection with a community or individual's past. For example, singing songs that are traditional to one's culture provides a tangible link to earlier members of one's community, even though the asset itself is intangible. This connection will create historical value. *Authenticity values* are generated when a cultural heritage asset is a genuine artifact of the culture; the value associated with this authenticity is distinct from the other cultural values that may be associated with the asset. Communities may derive *aesthetic values* from a cultural heritage asset due to its beauty or design or the placement of the heritage asset in the landscape, such as the façade of a cathedral or the arrangement of boulders at Stonehenge. Some cultural heritage assets may take on *symbolic* or even *iconic values,* such as the cedar tree in Lebanon that adorns their flag. The cedar tree is not just a tree for the Lebanese but rather a cultural symbol of strength and longevity. *Spiritual values* may be derived from cultural heritage assets such as cathedrals and churches whose sites invoke connection to one's spiritual identity or connectedness with other members of the same spiritual community. The aspects of culture that are defined by shared values and beliefs will generate *social values* for communities and provide a sense of connection with others in the community.

Other types of values may be assigned to cultural heritage assets as well. Borrowing from the literature in natural resource and environmental valuation, *bequest values* are those values that we hold for assets merely because we wish to be able to pass them on to our heirs. Cultural heritage assets may generate significant bequest values due to their historic, symbolic, and authenticity values.

Markets are not likely helpful in determining these cultural values for two reasons. The first reason is that many heritage assets are not exchanged in markets, and thus market prices do not exist that might serve to proxy or provide a fractional estimate of the total value of a cultural heritage asset. The second reason is that even when there is a market for cultural heritage asset (or an attribute of it), the market price will likely not reflect the cultural value assigned to that asset due to the nature of cultural heritage assets as public goods and/or those exhibiting positive externalities. As a result, nonmarket valuation techniques such as contingent valuation are frequently used in the economics of cultural heritage. Later in the chapter, we'll see that some methods will be incapable of measuring the cultural values of heritage sites, which has implications for the empirical work in the economics of cultural heritage.

The previous discussion has implied that economic values are likely less than the total value of a cultural heritage asset that includes its cultural values. The relative weight of economic and cultural values in total value is an empirical question, although it is likely to vary by type of cultural heritage asset. It may be that for many cultural heritage assets, the economic (market) value of a particular heritage asset is low, while the cultural heritage values are high. For example, an old industrial site may have great cultural value for its ability to tangibly depict the historical importance of a particular manufacturing technique or way of life, but it may have a very low property value.

One might ask whether it is appropriate to concern ourselves with the individual components of value, when the total value is what is effectively useful for the study of the economics of cultural heritage. Insofar as these values are unable to be represented in market prices, it is important to

catalog them so that we can create a more holistic picture of the overarching or total value for the cultural heritage asset in question. This will be helpful for managers of heritage assets and policy makers who are determining the policies that provide support for such assets.

It may also be that by deconstructing the value of a cultural heritage asset into its composite elements and querying consumers about them, we can learn more about the importance of that asset while also learning about how and why consumers formulate preferences. The estimation of willingness to pay (WTP), discussed more thoroughly in the next section, is presumed to be a proxy for value, and it is hypothesized that the factors that influence WTP include the presence or absence of the cultural values described above. While neoclassical economics has traditionally been uninterested in the preference formation process, instead presuming that we have a well-defined set of preferences a priori, it is clear that additional knowledge about consumer preference formation will benefit the discipline as a whole. Models of consumer behavior, for example, would be enriched by including models of preference formation.

### Cultural Heritage as a Capital Asset

Because cultural heritage is effectively a bundle of assets, the valuation of those assets is a core concern of the economics of cultural heritage. Learning how much citizens value historic monuments is one way of measuring the value of cultural heritage capital assets. Another is to ask how much they would be willing to pay to preserve those assets or to protect them from quality degradation. The answers to these questions are essential to understanding the amount of wealth that a nation has, how a citizen's quality of life is affected by the presence or absence of such assets, and the role of heritage assets in economic activities. In addition, a valuation of cultural heritage assets is an essential input into policy questions regarding the level of investment and regulation of cultural heritage assets.

As previously indicated, cultural heritage assets may be tangible or intangible, built or natural, permanent or ephemeral. This cultural heritage capital is thus uniquely differentiated from the physical capital, human capital, and natural capital that are typically studied by economists. However, there are similarities with other forms of capital, especially with regards to the decisions that are made that affect their quantity and quality. In theory, cultural heritage capital will continue to accumulate over time as a culture evolves, and we can encourage the accumulation of cultural capital to accelerate this growth. If this type of policy is pursued, the aggregate value of a region's cultural heritage assets will increase over time.

However, some forms of cultural heritage capital (monuments, architecture) must also be maintained to ensure that the asset does not deteriorate over time. Of course, it is likely that the condition of the heritage capital will determine some or all of the values that people hold for those assets, a testable hypothesis for researchers in the economics of cultural heritage. If a historic barn or church is left to disintegrate into the landscape, then eventually there may be no remnants of the site that induce cultural value. As a result, assessing the condition of heritage capital, as well as the change in that condition, is required just as it is essential to measure depreciation for other capital assets. The decisions regarding the accumulation, maintenance, and deterioration of cultural heritage capital assets are shared by individuals and government, as described in the previous section. Concerns about intergenerational equity are inherent in the management of cultural heritage assets since costs and benefits may not be evenly distributed across time or space.

## Applications and Empirical Evidence

### Valuation Studies

The value of cultural heritage can be estimated using several different methods. The appropriate choice of method will be determined by the type of cultural heritage asset that is to be valued such as whether it is fixed in a given geographic location, whether it draws visitors, whether it is tangible or intangible, and whether significant cultural values are believed to exist for the asset. This section will explore the diverse types of methods used to estimate the value of cultural heritage empirically and issues associated with this applied valuation work in the economics of cultural heritage.

The most frequently applied method for valuing cultural heritage has been the contingent valuation method (CVM). The CVM is a *nonmarket valuation technique,* a name given to the set of methodologies for valuing goods and services that are not exchanged in markets. Nonmarket valuation (NMV) techniques were developed to estimate the benefits associated with the attributes of the environment and natural resources that do not have market exchanges to determine price, such as clean air and water or recreation that is not marketed. The most common NMV technique, CVM, asks beneficiaries directly about their willingness to pay for a particular good or service. The method has been openly criticized because the method presumes a hypothetical, rather than actual, market and risks introducing biases without careful study design. The reliability of the CVM was thoroughly investigated by a National Oceanic and Atmospheric Administration (NOAA) panel, which concluded with a cautious stamp of approval for the use of CVM in natural resource damage assessment (NOAA, 1993). Since then, thousands of CVM studies have been used to estimate the value of natural resource and environmental amenities, due in part to the fact that it is a flexible method that can be applied to virtually any good or service. It is for this reason that CVM has been used in the estimation of cultural heritage values. CVM has been used to estimate many cultural heritage

assets, including museum sites, cathedrals, the medina in Fes (Morocco), and monasteries, among others. Because of its inherent flexibility, the CVM has also been frequently used to estimate the value of additional protection or preservation for cultural heritage assets. Navrud and Ready (2002) provide a sampling of such studies.

Another nonmarket valuation method that can be used to estimate some of the values of cultural heritage assets is the travel cost method. The travel cost method (TCM) uses recreational trip costs as a proxy for site value, and thus it is only a relevant method for those heritage sites that generate recreational visitation. Because the travel cost method presumes that nonvisitors have no value for the site, it is a less than perfect method for many heritage sites that will be valued by individuals who do not actually visit them. One of the first studies to use the travel cost method to value a cultural heritage site was Poor and Smith (2004), who estimated the value of historic St. Mary's City of Maryland.

As in the environmental and natural resource literature, studies of the economics of cultural heritage have combined the travel cost and contingent valuation methods. The advantage of combining the two methods is that actual price information revealed in travel costs can serve to mitigate the hypothetical nature of the contingent valuation exercise. Alberini and Longo (2006) combined travel cost and contingent behavior methods to study the value of cultural heritage sites in Armenia.

Choice modeling is another nonmarket valuation method that has been used to estimate the value of cultural heritage as an element in the overall valuation of a heritage site. In a choice modeling study, respondents are asked to simultaneously value the various attributes of a good or service by selecting from various bundles of characteristics for the asset in question. For example, respondents could be asked about different scenarios for a particular heritage site that are defined by varying levels of protection for cultural sites, differing levels of monetary contribution, and varying levels of access to the heritage asset. Choice modeling has been used to investigate the value of aboriginal cultural heritage sites (Rolfe & Windle, 2003) and the heritage values associated with farmland in western North Carolina (Mathews & Bonham, 2008).

Because some cultural heritage values are likely to be embedded in property values, the hedonic price method has also been applied to uncover the value of cultural heritage. The hedonic price method examines market prices for a good such as housing as a function of its component characteristics, including both housing characteristics (number of bedrooms and bathrooms, etc.) and other characteristics, including attributes such as air quality and proximity to amenities, recreation sites, or heritage assets. Rosato, Rotaris, Breil, and Zanatta (2008) use the hedonic method to explore whether housing prices in the Veneto region of Italy vary due to proximity to built heritage sites such as historical palaces, fortresses, and religious buildings; the variation in housing prices represents a value of cultural heritage.

While several methods can be used to estimate the value of cultural heritage, each of them is imperfect. The limitations of the travel cost method dictate that it will underestimate cultural values by assuming that only site visitors have value for them. The hedonic method can capture the component of cultural value that may be embedded in property values, but it is likely that many cultural values that we hold for heritage sites are accruing to individuals who do not own property proximate to them, and thus the hedonic method, too, will provide an underestimate of the total value of cultural heritage. Choice models more closely mimic market transactions than the contingent valuation method, but they are challenging to design due to their complex nature and may be confusing for respondents to complete. The contingent valuation method is the only method that is likely to be able to capture the full value that we have for cultural heritage assets (economic + cultural values), which is likely why we have seen relatively more CVM conducted than any other method. Additional methodological advances in the valuation of cultural heritage, perhaps by strengthening the field's connection with environmental and natural resource economics, would help resolve some of the issues noted here.

## The Economic Impact of Cultural Heritage

Most studies estimating the value of cultural heritage have investigated built sites such as monuments and historic buildings. Because these historic sites can attract visitors, and because those visitors expend scarce dollars to experience the heritage assets at the site that benefit communities in the form of sales and tax revenue, there have been many studies estimating their economic impact. The economic impact studies provide what might be considered a lower bound value for the sites since they cannot provide an accounting of the cultural values ascribed to the sites. However, it is likely that the cultural values that consumers hold for these sites are a factor in the demand for visits and thus an important underlying preference.

Until recently, most studies of the economics of cultural heritage were studies estimating the economic impact of tangible or built heritage assets such as monuments and architecture. A vast majority of these studies have been done outside the United States, in both Europe and developing countries. This may be because the portfolio of cultural heritage assets is richer for countries with longer histories than the United States or because the widespread public interest in those heritage assets provides a rationale for their study.

The role that cultural heritage plays in attracting tourists has led to several tourism studies of cultural heritage sites. These studies have implications for both their economic impact and the management of the sites themselves. Cuccia and Cellini (2007) examined the preferences of tourists visiting Scicli, a Sicilian town known for its baroque heritage, and found that cultural heritage was not among the most important reasons for visitors making

their trip. Other studies have examined the behaviors of tourists at cultural heritage sites with the aim of providing recommendations for tourism management (de Menezes, Moniz, & Vieira, 2008; Kolar & Zabkar, 2007) and whether or not World Heritage Site listing increases tourism (Tisdell & Wilson, 2002).

In addition to studies estimating the economic impact of visitation to heritage sites, the economic impact of construction and maintenance expenditures for cultural heritage assets has been investigated. In the European Union, for example, it has been estimated that 50% of all construction activity is related to building restoration work (Cassar, 2000). Thus, the protection of cultural heritage assets can provide a significant contribution to a region's economy.

### Cultural Heritage and Economic Development

The more general question of the relationship between cultural heritage and economic development is, as one might expect, also a concern for the economics of cultural heritage. For example, Murillo Viu, Romani Fernandez, and Surinach Caralt (2008) investigated the impact of the Alhambra on the economy of Grenada, Spain. Additional studies of heritage sites' role in economic development have been predominantly focused on developing countries, due in part to the role that institutions such as the World Bank have had in developing place-based economic development strategies that include protection of cultural heritage sites.

## Policy Implications

### Government Provision of Cultural Heritage

Because the public goods nature of cultural heritage will lead to the underprovision of heritage by markets, there is a motivation for the government provision of cultural heritage. Several studies have examined whether public support exists for additional government activities to protect cultural heritage. The outright, direct provision of cultural heritage is frequently pursued by governments as evidenced by the Smithsonian Institution in the United States and in national parks, monuments, and historic sites across the globe. An infrequently pursued but perfectly applicable empirical investigation for the economics of cultural heritage is cost-effectiveness analysis of these government investments in cultural heritage.

Government intervention is also prescribed if the provision of cultural heritage assets has spillover benefits (positive externalities) since the market will also tend to provide fewer heritage assets than would be optimal for society in this case. Governments frequently use subsidies and tax policy to promote private provision and/or preservation of heritage assets. In addition, regulation is frequently used by governments, despite the fact that it is often the economist's least favorite tool. Historic districts that restrict design and construction and even color of homes are commonly used to promote or ensure the provision of cultural heritage by private individuals.

Another interesting question that has been identified for the economics of cultural heritage is the question of who benefits from government investment in cultural heritage and who pays (Throsby, 1997). Social benefit-cost analyses of cultural heritage projects have not been widespread, although it is hypothesized that the benefits are not as well distributed as costs. While we might expect that everyone benefits equally from the existence of cultural heritage, since individuals with higher education and income levels are more likely to visit or otherwise consume certain types of cultural heritage, it is not clear that the distribution of costs and benefits is coincident. Additional studies on the distribution of the costs and benefits of providing cultural heritage will greatly enrich the field.

## Emerging Opportunities in the Economics of Cultural Heritage

The economics of cultural heritage has up to now positioned itself as many other applied fields in economics, where conventional methodologies (economic impact analysis, nonmarket valuation techniques, welfare analysis) are applied to new settings. While the value of cultural heritage will likely continue to reign as the primary investigative theme for some time to come, several emerging opportunities indicate the field may be approaching an adolescence of sorts. If these opportunities are pursued, the economics of cultural heritage may very well push the boundaries of economics into exciting new territory in the twenty-first century.

One opportunity for researchers that has yet to be realized is to study in depth the nature of cultural heritage as cultural capital. In particular, one interesting question would be to investigate the change in the value of heritage assets over time. In theory, if we construe cultural heritage as another form of capital asset, then the value of these assets, if maintained, should rise over time. Exploring the rate of appreciation of these assets, especially with the aim of making comparisons with other assets in which the government may invest, is a fruitful direction for future research that could yield significant implications for government policy. This line of inquiry requires both a benchmark valuation for cultural heritage capital and a commitment to regular investigation of a specific asset, and thus it will be more likely that we would see this type of research conducted with tangible heritage assets that are easier to define and monitor than intangible heritage assets.

Another opportunity for the economics of cultural heritage is to strengthen its link with natural resource and environmental economics by increasing the recognition and

importance of natural assets in defining cultural heritage. To date, very few economic studies have been conducted that define cultural heritage via a region's natural assets. One such study investigated the value of Armenia's Lake Sevan as a unique symbol of Armenian cultural heritage (Laplante, Meisner, & Wang, 2005). However, there are many cases around the globe where natural assets are important components of a region's cultural heritage that would provide interesting and significant case studies for the economics of cultural heritage. One example exists in Lebanon, where the cedar tree, which adorns the Lebanese flag, is an important part of Lebanese cultural heritage. These trees, some as much as 3,000 years old, have been nominated as one of the Seven Wonders of the World and are currently under threat due to global climate change. Investigating the value of protecting these cultural heritage assets, as well as other forms of heritage that exist in natural resources, would both advance the work of the economics of cultural heritage and environmental valuation methodologies.

Along these lines, it would be interesting to know how the global citizenry values iconic cultural heritage assets such as UNESCO-designated World Heritage Sites. To date, most studies have focused on estimating the values that regional citizens or visitors hold for their proximate cultural heritage assets. Learning the values held by the global population could assist in designing effective policies for their long-term survival.

A related opportunity is to more intentionally situate the study of the economics of cultural heritage in a given geographic space. Although much of our heritage is specifically embedded in a particular geographic location or place, very few studies have intentionally incorporated the use of geographic information systems and other spatial methods to better understand the value of cultural heritage assets. One example is the interdisciplinary Farmland Values Project that investigates the values that four communities in western North Carolina have for farmland. One of the methodologies used in the study had participants describe the places that were culturally important to them using a widely and freely available mapping software, GoogleEarth, and rate places they had located on the map for their heritage and other attributes. The exercise yielded a community-defined map of culturally important places that is intentionally embedded in place. Additional work in the economics of cultural heritage could serve to strengthen the methodologies for both collecting and spatially organizing cultural heritage valuation data and work toward building a bridge between economics, geography, and other fields such as ethnobotany and cultural anthropology that have more intentionally place-based modes of inquiry.

At the outset of the chapter, the definition of cultural heritage included intangible elements such as customs, yet no studies have investigated the value of cultural customs, ways of life, or other intangible aspects of cultural heritage. Investigations in this area would be an excellent bridge between the traditionally quantitative methods of economics and the more qualitative methods of anthropology and sociology. Thus, in addition to broadening the repertoire of the economics of cultural heritage, research in this area could help to push the boundaries of the economic discipline by gathering information on how and why consumers formulate preferences for intangible goods and services.

## Conclusion

The economics of cultural heritage is an emerging subfield in economics that has the ability to both serve academic audiences outside of economics (sociology, anthropology, history, political science) and push the boundaries of the economics discipline as a whole. The development of the field has been closely linked with cultural economics and shares with it the importance of incorporating knowledge from disciplines that traditionally have not been rigorously studied by economists, including history, architecture, and the arts. The future of the economics of cultural heritage looks bright as there are numerous opportunities for the field to make significant advances in the valuation of cultural heritage and for pushing the boundaries of economics. Some of these include a more intentional incorporation of interdisciplinary methods and more complex studies to evaluate the full set of values that we hold for our cultural heritage, including the truly intangible elements.

## References and Further Readings

Alberini, A., & Longo, A. (2006). Combining the travel cost and contingent behavior methods to value cultural heritage sites: Evidence from Armenia. *Journal of Cultural Economics, 30,* 287–304.

Cassar, M. (2000, September). *Value of preventive conservation.* Keynote lecture at the European Preventive Conservation Strategy Meeting, Institute of Art and Design, Vantaa, Finland. Retrieved January 3, 2008, from http://www.ucl.ac.uk/sustainableheritage/keynote.htm

Cuccia, T., & Cellini, R. (2007). Is cultural heritage really important for tourists? A contingent rating study. *Applied Economics, 39,* 261–271.

de Menezes, A. G., Moniz, A., & Vieira, J. C. (2008). The determinants of length of stay of tourists in the Azores. *Tourism Economics, 14,* 205–222.

Ginsburgh, V., & Throsby, D. (Eds.). (2006). *Handbook of the economics of art and culture* (Vol. 1). Amsterdam: Elsevier, North-Holland.

Hutter, M., & Rizzo, I. (Eds.). (1997). *Economic perspectives on cultural heritage.* New York: St. Martin's.

Kolar, T., & Zabkar, V. (2007). The meaning of tourists' authentic experiences for the marketing of cultural heritage sites. *Economic and Business Review, 9,* 235–256.

Laplante, B., Meisner, C., & Wang, H. (2005). *Environment as cultural heritage: The Armenian diaspora's willingness-to-pay to protect Armenia's Lake Sevan* (Policy Research Paper, WPS 3520). Washington, DC: The World Bank.

Mathews, L. G., & Bonham, J. (2008, June). *Pricing the multiple functions of agricultural lands: Lessons learned from the Farmland Values Project.* Paper presented at the annual meeting of the Western Agricultural Economics Association, Big Sky, MT.

Murillo Viu, J., Romani Fernandez, J., & Surinach Caralt, J. (2008). The impact of heritage tourism on an urban economy: The case of Granada and the Alhambra. *Tourism Economics, 14,* 361–376.

National Oceanic and Atmospheric Administration (NOAA). (1993). Report of the NOAA panel on contingent valuation. *Federal Register, 58,* 4602–4614.

Navrud, S., & Ready, R. C. (Eds.). (2002). *Valuing cultural heritage: Applying environmental valuation techniques to historic buildings, monuments and artifacts.* Cheltenham, UK: Edward Elgar.

Poor, P. J., & Smith, J. M. (2004). Travel cost analysis of a cultural heritage site: The case of historic St. Mary's City of Maryland. *Journal of Cultural Economics, 28,* 217–229.

Richards, G. (Ed.). (2007). *Cultural tourism: Global and local perspectives.* New York: Haworth.

Rizzo, I., & Towse, R. (Eds.). (2002). *The economics of heritage: A study in the political economy of culture in Sicily.* Cheltenham, UK: Edward Elgar.

Rolfe, J., & Windle, J. (2003). Valuing the protection of aboriginal cultural heritage sites. *Economic Record, 79,* S85–S95.

Rosato, P., Rotaris, L., Breil, M., & Zanatta, V. (2008). *Do we care about built heritage? The empirical evidence based on the Veneto House Market* (Working Paper 2008.64). Milan, Italy: Fondazione Eni Enrico Mattei.

Throsby, D. (1997). Seven questions in the economics of cultural heritage. In M. Hutter & I. Rizzo (Eds.), *Economic perspectives on cultural heritage* (pp. 13–30). Basingstoke, UK: Macmillan.

Throsby, D. (2001). *Economics and culture.* Cambridge, UK: Cambridge University Press.

Tisdell, C., & Wilson, C. (2002). World heritage listing of Australian natural sites: Tourism stimulus and its economic value. *Economic Analysis and Policy, 32,* 27–49.

Towse, R. (Ed.). (2003). *A handbook of cultural economics.* Cheltenham, UK: Edward Elgar.

Towse, R. (Ed.). (2007). *Recent developments in cultural economics.* Cheltenham, UK: Edward Elgar.

Vaughn, D. R. (1984). The cultural heritage: An approach to analyzing income and employment effects. *Journal of Cultural Economics, 8,* 1–36.

# 81

# Media Economics

Gillian Doyle

*Centre for Cultural Policy Research, University of Glasgow*

Media economics combines the study of media with economics. The term *media* is usually interpreted broadly and includes sectors such as television or radio broadcasting plus newspaper, magazine, or online publishing; communications infrastructure provision; and also production of digital and other forms of media content. Media economics is concerned with unravelling the various forces that direct and constrain choices made by producers and suppliers of media. It is an area of scholarship that has expanded and flourished in departments of economics, business, and media studies over the past two decades.

A number of reasons explain why media economics has advanced quite significantly in popularity and status over recent years. The increasing relevance of economics has been underlined by the so-called digital revolution and its effect in reshaping media businesses while, at the same time, accelerating related processes of convergence and globalization. Deregulation of national media industries is another major trend that has shifted attention on the part of media policy makers and also academics from political toward economic issues and questions. So although media economics—the application of economics theories and concepts to all aspects of media—is still at a relatively early stage of development as a subject area, its importance for industry, policy makers, and scholars is increasingly apparent.

The earliest studies of economics of mass media can be traced back to the 1950s, and these looked at competition among newspapers in the United States. Competition and concentrations of ownership are still key and constant themes within media economics, notwithstanding the many shifts and changes that have redrawn the competitive landscape over time. Other early work that marked out economics of media as being a distinctive field includes studies of competitive programming strategies (i.e., of the different program content strategies used by competing broadcasters).

Some landmark studies in media economics owe their existence to the needs of policy makers who have asked for work on, for example, competitive conditions within specific sectors of industry or questions around market access or issues such as spectrum pricing. A very good example is the Peacock (1986) report, commissioned by the U.K. government ahead of the 1990 Broadcasting Act. As the first systematic economic assessment of the U.K. television industry, this report was to have seminal influence over subsequent broadcasting policy in Britain. More recently, a wave of interest, initially sparked by Richard Florida's (2002) work on urban economics, has fuelled demand for work by economists on "creative industries" (which include media content production). Studies in this area (see, e.g., Hutton, O'Keefe, Schneider, Andari, & Bakhshi, 2007) are frequently concerned with the capacity for creative industries to drive forward growth in the wider economy.

The origins and approaches evident in economic studies of the media are varied. Some work has been theoretical, seeking to build on approaches within mainstream economics and, occasionally, to develop specialized models that take account of the special contingencies of the media industry. Much work so far in this subject area has tended to be in the applied tradition, looking at specific markets and firms under specific circumstances.

Generally, the broad concern is how best to organize the resources available for provision of mass media. Economists specializing in media economics have explored whether firms are producing the right sorts of goods and services and whether they are being produced efficiently.

Some (frequently drawing on an industrial organization approach) have examined the association between the markets media firms operate in and their strategies or their performance or their output. Another common concern, especially in work on broadcasting, has been what role the state should play in ensuring that the organization and supply of media output matches societal needs.

While research within the tradition of media economics has spanned across all aspects of all media sectors, including film, television, radio, newspapers, magazines, and the Internet, it is worth noting that an overlap exists between this and the related area of "cultural" economics. Cultural economics has a wider ambit; covering arts and heritage as well as media and cultural economics has developed as a separate field with its own concerns (such as subsidies for the arts). But there is some common ground, for instance, concerning the economics of creativity or optimal levels of copyright protection. Research into international trade in films, for example, can be regarded as equally at home in either of these fields.

As well as what might be termed *mainstream economic research into mass media,* the field of media economics is also strongly populated by work that emerges from political economy traditions. The "critical" political economy approach links sociopolitical with economic analysis. It adopts a more normative approach to the analysis of economic actors and processes, rather than focusing simply on, say, questions of efficiency. A number of influential thinkers in media and communications, such as Bagdikian, Garnham, and McChesney, have emerged from the critical political economy tradition. One such is Douglas Gomery (1993), who points out that "studying the economics of mass communication as though one were simply trying to make toaster companies run leaner and meaner is far too narrow a perspective" (p. 198).

So, media economics is a diverse and lively area of scholarship that draws on many different sorts of approaches. Those coming new to the subject will find a number of textbooks on hand to provide a basic understanding of economic concepts and issues in the context of media. The emergence of such books was traced recently by Robert G. Picard (2006), one of the leading figures in this subject. The first textbook appeared in French back in 1978 (Toussaint-Desmoulins), followed by a Spanish-language text in the mid-1980s (López, 1985) and, later, the first German text (Bruck, 1993). Picard himself wrote the first introductory textbook in media economics in the English language (1989) and, in surveying later contributions from the United States, he (2006, p. 21) draws attention to textbooks by Alexander, Owers, and Carveth (1993); Albarran (1996); and Owen and Wildman (1992). Also highlighted is a textbook by a U.K.-based author (Doyle, 2002) that blends traditional economics along with political economy perspectives.

Helpful though textbooks are, the depth and diversity of the field can only be fully appreciated through acquaintance with the growing range of scholarly books, journal articles, monographs, research reports, and studies that focus on economic aspects of media. A rich and diverse body of literature has emerged over the years from a variety of sources, and as the subject grows, media economics continues to expand both in its ambitions and popularity.

This chapter introduces some key themes that are characteristic of media economics and have shaped the development of the field. This survey is not exhaustive, and although the discussion is broken into sections, in reality there are numerous overlaps and interconnections between topics and concerns central to this area of scholarship. In highlighting core issues that students working in the area of media economics are likely to encounter, the main aim here is to provide an introductory overview and a sense of what is special and interesting about this particular subfield within economics.

## Economics of Media Is Different

One of the main attractions as well as a key challenge of carrying out work on economics of media stems from the fact that media are a bit "different" from other commodities. It is sometimes said that media operate in dual-product markets—generating not only media output (i.e., content or messages) but also audiences (i.e., the viewers or readers who are attracted by the output) (Picard, 1989, pp. 17–19). The peculiarities of media as a commodity relate mostly to first sort of product: media content.

Collins, Garnham, and Locksley (1988) were pioneers in explaining the economic peculiarities of the broadcasting commodity. These authors flag up a similarity between broadcast output (e.g., a program broadcast on television) and other cultural goods insofar as "the essential quality from which their use-value derives is immaterial" (p. 7). Many cultural goods share the common characteristic that their value for consumers is tied up in the messages or meanings they convey, rather than with the material carrier of that information (the radio spectrum, CD, or the digital file, etc). Because messages or meanings are intangible, media content is not "consumable" in the purest sense of this term (Albarran, 1996, p. 28). Because of the "public good" characteristic of not being used up or not being destroyed in the act of consumption, broadcast material exhibits the peculiarity that it can be supplied over and over again at no extra cost. If one person watches a TV broadcast or listens to a song, it does not diminish anyone else's opportunity to view or listen. In this respect, media seem to defy one of the very basic premises on which the laws of economics are based—scarcity.

The various insights offered by Collins et al. (1988) about, for example, the nonrivalrous and nonexcludable nature of broadcast output were an important early landmark in the development of thinking about how the economic characteristics of mass media differ from other

industries. But later work by Richard Caves (2000) again highlighted the requirement for any understanding of the economics of creative industries, of which media are a part, to be based around an appreciation of the peculiarities of that sector. His influential work on "art and commerce" (Caves, 2000) applies economic analysis to the special characteristics of creative activities (e.g., uncertainty of demand, incentives and motivations guiding artistic and creative "talent") and, in so doing, explores various aspects of the organization and behavior of creative industries.

Students and researchers cannot escape the challenges that derive from the distinctive nature of their area of enquiry. One such, in media economics, is the difficulty of measuring or evaluating the impacts that arise from a decision to allocate resources in one fashion rather than another. Communicating with mass audiences, as an economic activity, is inextricably tied up with welfare impacts. And media economics seeks to play a role in showing how to minimize the welfare losses associated with any policy choices surrounding media provision. But, as prominent economist Alan Peacock (1989, pp. 3–4) observed some years ago, welfare impacts are and still remain very difficult to measure convincingly.

Another problem is that, whereas notions of economic "efficiency" and assessments of whether efficiency is being achieved depend on clarity about objectives, the circumstances surrounding cultural provision often militate against such clarity. The perceived objectives associated with media provision are varied and at times contradictory, with some organizations operating in the nonmarket sector (Doyle, 2002, p. 11). So, when it comes to analyzing media, the application of all-embracing models based in conventional economic theory often proves inadequate. Thus, as many have observed, an ongoing concern for economists specializing in media is to build on and develop suitable and coherent overarching theories and paradigms for the study of this as a particular subject area (Fu, 2003; Gomery, 1993; Lacy & Bauer, 2006; Wirth & Bloch, 1995).

Other special challenges that media economists are faced with stem from, for example, the uncertainties, risks, and irrationalities associated with producing creative output or from seeking to analyze production, distribution, and consumption in an ever-changing technological environment for mass media (Doyle & Frith, 2006). The business of supplying ideas and information and entertainment to mass audiences is different from supplying other ordinary commodities such as baked beans, but the complexities that go along with this are central to the legitimacy as well as to the unique appeal of media economics as a distinctive subject area.

## Audiences and Advertising

The business of media is about supplying audiences as well as forms of content, and indeed, many mass media are supported largely through advertising revenue. So, not surprisingly, work on audiences and their behavior and around advertising has featured strongly in media economics research and scholarship to date. The vast influence that patterns of advertising activity exert over the fortunes of the media industry has been underlined by the 2008 banking crisis and associated economic recession where a diminution in expenditure on advertising in newspapers, magazines, and broadcast channels has prompted wide-scale closures and job losses across the media in the United States and Europe. Work on the economics of advertising has addressed questions around the relationship between economic wealth and advertising, cyclicality in advertising, the economic role played by advertising, and the impact it exerts over competitive market structures and over consumer decision making (Chiplin & Sturgess, 1981; Schmalensee, 1972; Van der Wurff, Bakker, & Picard, 2008).

Audiences are another focus of interest. A number of studies have examined the nature of audiences (or of access to audiences) as a commodity, audience ratings, and how demand among advertisers for audience access is converted into revenue streams by media enterprises (Napoli, 2003; Webster, Phelan, & Lichty, 2000; Wildman, 2003). Audience fragmentation, although present as an issue in early work about television audiences (Barwise & Ehrenberg, 1989), has risen to greater prominence in recent studies. As Picard (2002) suggests, fragmentation of mass audiences is "the inevitable and unstoppable consequence of increasing the channels available to audiences" (p. 109). In a recent study, Webster (2005) concludes that, despite ongoing fragmentation, levels of polarization among U.S. television audiences are modest so far. Nonetheless, the continued migration of audiences toward digital platforms and associated processes of fragmentation raises important questions for future work in media economics (and in media sociology too).

## Media Firms, Markets, and Competition

Although some studies of the relationship between the economy and advertising are macroeconomic, most work by economists interested in media fits within the category of microeconomics. A central focus of interest is firms and how they produce and supply media and also the markets in which media organizations operate and levels of competition.

The concept of a media firm covers many different sorts of actors, but what they all have in common is an involvement somehow in producing, packaging, or distributing media content. Of course, all media firms are not commercial organizations. The prevalence, initially within broadcasting but now across digital platforms too, of nonmarket organizations devoted to providing public service content means that standard assumptions about profit maximization that are central to the theory of firms become questionable in the context of media. Another complicating factor is that ownership of media such as national newspapers is sometimes motivated by concerns that have

little to do with economics, such as, especially, the pursuit of political influence. So firms are important within media economics research, but standard economic theories about the behavior of firms have their limitations in this context.

Be that as it may, the industrial organization (IO) model, which is based on the theory of firms, offers an analytical framework that has frequently proven useful to economists working on media firms or industries (Hoskins, McFadyen, & Finn, 2004, pp. 144–156). The IO model (and associated structure-conduct-performance or SCP paradigm) suggests that the competitive market structure in which firms operate will, in turn, affect how they behave and subsequently their performance. Although some doubts have been cast on the causal links of the SCP in recent years, it remains that many media economists have profited from the broad insights offered by IO theory about how, in practice, media firms behave under different market structures and circumstances.

Approaches toward analyzing media firms and markets sometimes take as their starting point the concept of the "value chain" approach first developed by Michael Porter (1998). The media industry can be broken up into a number of broad stages, starting first with production or creation of content (which usually, though not always, brings initial entitlement ownership of intellectual property), then assembling content into services and products (e.g., a newspaper or television channel), and finally distribution or sale to customers. Definitions of markets and sectors in media economics studies are often implicitly or explicitly informed by this conceptual framework. All of the stages in the vertical supply chain are interdependent, and this has important implications for the kinds of competitive and corporate strategies media firms will pursue.

A notable feature of the economics of media is that firms in this sector tend to enjoy increasing marginal returns as their output—or, more properly, *consumption* of their output—expands. The prevalence of economies of scale is strongly characteristic of media industries, and the explanation for this lies in the "public good" nature of the product and how it is consumed. Because the cost of producing a newspaper or supplying a television service is relatively unaffected by how many people choose to consume that output, it follows that these activities will enjoy increasing returns to scale. Plentiful studies exist that confirm the tendency toward high initial production costs in the media sector accompanied by low marginal reproduction costs. The cost of producing a feature film, a music album, or a television program is not affected by the number of people who are going to watch or listen to it. "First-copy" production costs are usually high, but then marginal reproduction or distribution costs are low and, for some media suppliers, zero.

Another important feature is the availability of economies of scope. Economies of scope are generally defined as the savings available to firms from multiproduct production or distribution. In the context of media, economies of scope are common, again because of the public good nature of media output and the fact that a product created for one market can, at little or no extra cost, be reformatted and sold through another. Because the value of media output is contained in messages that are intangible and therefore do not get used up or "consumed" in the traditional sense of the word, the product is still available to the supplier after it has been sold to one set of consumers to then sell over and over again. The reformatting of a product intended for one audience into a "new" one suitable to facilitate additional consumption (e.g., the repackaging of a celebrity interview into a television news package, a documentary, a radio transmission, etc.) releases savings for the firm and therefore generates economies of scope (Doyle, 2002, pp. 4–15).

In any industry where economies of scale and scope are present, firms will be strongly motivated to engage in strategies of expansion and diversification that capitalize on these features. This is certainly true of media. Concentrations of ownership within and across sectors of the media are a highly prevalent feature of the industry. As a result, many scholars working in the area of media economics have taken an interest in questions about sustaining competition and diversity, measurement of concentration levels, and more generally around how strategies of expansion affect the operation, efficiency, and output of media suppliers.

The link between ownership patterns and diversity within the output offered by media firms is one area of enduring interest. Theories of program choice, the early versions of which are reviewed in Owen and Wildman (1992), are concerned with under what conditions—including the number of competing channels and ownership of broadcast channels—the marketplace will offer similar as opposed to different sorts of programming, or cheap as opposed to expensively produced programs, and so on. The connection between diversity of ownership and output has also been studied in the context of the music industry and the film industry.

Some studies (e.g., Albarran & Dimmick, 1996) have focused on defining and measuring concentration levels within media markets using either the Herfindahl-Hirschman Index (HHI) or a "concentration ratio" such as CR4 or CR8. Others are more concerned with analyzing the economic motivations that underlie strategies of expansion and diversification by media firms. Sánchez-Tabernero and Carvajal (2002) and others have analyzed advantages and also risks associated with a variety of growth strategies, including horizontal, multimedia (cross-sectoral), and international expansion.

A great deal of work in media economics has concerned itself with changing market structures and boundaries within the media. Economics provides a basic theoretical framework for analyzing markets based on the clearly defined structures of perfect competition, monopolistic competition, oligopoly, and monopoly. In practice, many

media firms have tended to operate in markets whose contours are strongly influenced by technological factors, state regulations, or both. In addition, most media have tended to operate in very specific geographic markets and to be closely linked to those markets by the nature of their product and through relationships with advertisers. These factors have restrained levels of competition in the past. But times are changing, and this is reflected in much more fluid boundaries and competitive market structures.

Whereas, at the outset of broadcasting, market access was constrained by spectrum limitations, the structure of the television and radio industries has been transformed by the arrival and growth of new forms of delivery for television such as cable, satellite, and digital platforms. Many studies over the years have focused on the effect of channel proliferation and additional competition, as well as increased sectoral overlap between broadcasting and other forms of communications provision (Picard, 2006). Likewise, major changes in the economic organization of print media industries and their impact on production costs, market access, and levels of competition are frequently the subject of interest in media economics texts and studies.

Not only have the avenues for distribution of media been expanding, but also changes in technology and in state policies have opened up national broadcasting systems and contributed to a growing trend toward internationalization of operations by media companies. The process of globalization of media has been propelled forwards by digital technologies (Goff, 2006) and the growth of the Internet, which has created a major impetus in the direction of global interconnectedness. Much recent research work has addressed strategic responses on the part of firms both to common trends generally affecting the media environment, such as globalization and convergence (Kung, 2008), and to changes that are very specific to individual media markets, such as internationalization of Norwegian newspapers (Helgesen, 2002), deregulation of broadcasting in Finland (Brown, 2003), or the role of domestic quotas in the success of Korean films (Lee & Bae, 2004).

## Business Strategies

Media economics concerns itself with a wide range of strategies and behaviors that reflect the distinctive features and circumstances of this industry. Hoskins, McFadyen, and Finn (1997) have examined some key economic and managerial challenges facing firms in the television and film production industries, for example, the need to ensure that creative and business inputs function effectively alongside each other—an issue that has also been tackled in some depth by Caves (2000). Hoskins et al. explain how risks and uncertainties associated with producing high-cost audiovisual output are offset, for example, by the use of sequels and series that build on successful formats and through the "star" system, which helps to build brand loyalty among audiences and therefore promote higher and more stable revenue streams.

Strategies of risk spreading are important in media because of uncertainty surrounding the success of any new product. Production is expensive, and while market research may prove helpful, these are essentially hit-or-miss businesses. The factors that determine whether films, books, and music albums will prove popular (including fads, fashions, and the unexpected emergence of "star" talent) are difficult to predict. So strategies that counteract or mitigate risk are essential.

In television and radio, the fact that what is transmitted on any single channel is usually a whole range of products (a full schedule of programs) allows for some of the risks inherent to broadcasting to be reduced (Blumler & Nossitor, 1991, pp. 12–13). Control over a range of products greatly increases a broadcaster's chances of making a hit with audience tastes and therefore covering the cost of producing the whole schedule or "portfolio" or programs. In the twenty-first century, digital compression techniques and more channel capacity have extended the opportunities for broadcasters to engage in portfolio strategies because, as well as offering variety within the schedule of an individual channel, many television companies have become multichannel owners offering variety across a range of related services (MTV1, MTV2, etc.).

The success of the Hollywood majors in counteracting risk and dominating international trade in feature films has been of enduring interest for scholars working in media economics, including De Vany (2004), Hoskins et al. (1997), Steemers (2004), and Waterman (2003). The key to risk reduction for Hollywood producers is again to be found in control over distribution and the ability to supply audiences with a range of product. The ability to support and replenish a large portfolio of output is dependent on being able to fully exploit new and old hits but, as with the music industry, this is now potentially under threat from illegal copying.

Many of the strategies for economic advancement adopted by media firms are based on sharing content and therefore exploiting intellectual property assets as fully as possible. "Networking" is a good example. In broadcasting, a network is an arrangement whereby a number of local or regional stations are linked together for purposes of creating or exploiting mutual economic benefits (Owen & Wildman, 1992, p. 206). Usually the main benefit is economies of scale in programming. Because they are based in different localities, local or regional stations that form part of a network can successfully share a similar or identical schedule of programs, thereby reducing per-viewer costs by spreading the cost of producing that service across a much bigger audience than would otherwise be possible.

A similar sort of logic is at work in, for example, the affiliations or networks of international publishing partners

that may be involved (e.g., under franchise agreements) in publishing several different international versions of the same magazine title (Doyle, 2006). The ability to share content and images across the network means that the cost of originating copy material can be spread across a much wider readership, and each partner benefits from access to more costly elements of content (e.g., celebrity interviews) than could be afforded if the magazine were a stand-alone operation. Aside from reaping economies on content, being part of a network may also confer benefits in terms of shared deals on advertising sales.

The translation or reformatting of content from one media platform to another makes increasing economic sense in the context of globalization and digitization, especially so in times of economic recession when revenues are under pressure. This process, referred to by Murray (2005, p. 420) as "content streaming," involves the coordinated distribution of strongly branded content across multiple delivery formats. The aim is to reap economies not by using content that appeals to a single mass audience but rather through building and leveraging brand loyalties among specific target audience segments.

For media content suppliers, profit maximization depends on the full and effective exploitation of intellectual property rights across all available audiences. So a crucial concept in the economics of supplying media content is "windowing" (Owen & Wildman, 1992). This refers to maximizing the exploitation of content assets by regarding primary, secondary, and tertiary audiences (e.g., on free vs. pay channels) as "windows" and by selling your products not only through as many windows or avenues as possible but also in the order that yields the maximum possible return. The idea behind windowing is to carefully arrange the timing and sequence of releases so as to maximize the profit that can be extracted from the whole process, taking into account factors such as audience size, profit margin per head, risks of piracy, and so on.

When Owen and Wildman (1992) explained the practice of windowing, it was in the context of releasing programs via pay and free television and video outlets in the United States and overseas markets. Approaches toward exploiting content have evolved considerably since then, with many media operators now adopting "blended" distribution strategies and operating a "360-degree" approach to commissioning (Pennington & Parker, 2008). In other words, product is created with the intention of selling it across numerous different delivery platforms, not just television but also mobile and Internet. In an increasingly cross-platform or converged production context, the need to devise and execute strategies for effective and full economic exploitation of intellectual property assets has become more pressing and more complex. So windowing remains an important theme in media economics, with potential to offer useful theoretical and practical insights for all media content suppliers.

## Media Economics and Public Policy

It is usually assumed in economics that free markets will work better to allocate resources than centralized decision making by government, but intervention is sometimes called for to counteract deficiencies arising from the free operation of markets. This might be, for instance, because a "market failure" has occurred. A concern that is often at the root of work carried out in media economics is what role the state should play in ensuring that the organization and supply of media output matches societal needs. Which sorts of policies and what forms of intervention and regulation are needed to correct market failures and/or improve the allocation and usage of resources devoted to media provision?

So generally speaking, the most important economic reasons why state intervention might be needed are because of a need to address market failures, deal with the problem of "externalities," or restrict or counteract the use of monopoly power. But it is worth noting that governments can and do very often intervene in media markets for noneconomic reasons too. Because of the sociocultural and political influence that accompanies the ability to communicate with mass audiences, media and communications tend to be much more heavily regulated than other areas of economic activity, with special provisions covering, for example, protection of minors, balance and impartiality, and so on.

Broadcasting—still the largest sector of the media in economic terms—is, as evident from much of the literature of media economics, seen as especially prone to market failure. An example of failure is that broadcasting would not have taken place at all in the first place if left up to profit-seeking firms reliant on the conventional mechanism of market funding (i.e., consumer payments) because at the emergence stage in radio and television, there was no way to identify and/or charge listeners and viewers.

But many failures stem partly from the public good characteristics of the broadcasting commodity already mentioned. With any good or service that is "nonexcludable" (i.e., you cannot exclude those that do not want to pay) and where customers do not have exclusive rights to consume the good in question, free rider problems are virtually inevitable. Being a public good, broadcast output also has the characteristic of being "nonexhaustible"—typically, there are zero marginal costs in supplying the service to one extra viewer. Thus, it can be argued that "restricting the viewing of programmes that, once produced, could be made available to everyone at no extra cost, leads to inefficiency and welfare loss" (Davies, 1999, p. 203). Another cause of market failure relates to the problem of asymmetric information. Graham and Davies (1997) summarize this by explaining that "people do not know what they are 'buying' until they have experienced it, yet once they have experienced it they no longer need to buy it!" (p. 19).

Another source of market failure is externalities. These are external effects imposed on third parties when the internal or private costs to a firm of engaging in a certain activity (pollution, for example) are out of step with costs that have to be borne by society as a whole. Broadcasting can have negative external affects when, for example, provision of violent content imposes a cost on society (through increasing fear of violence among viewers). Because the cost to society is not borne by the broadcaster, there arises a market failure in that broadcasters may well devote more resources to providing popular programs with negative external effects than is socially optimal. And this needs to be corrected through some form of public policy intervention.

Externalities in the media are by no means always negative. It is generally recognized that some forms of media content confer positive externalities (e.g., documentaries, educational and cultural output) and, equally, that such output may be undersupplied under free-market conditions.

Although advances in technology have gone some way toward correcting initial causes of market failure (e.g., absence of mechanisms for direct payments), there are still a number of ways in which it might be argued that when it comes to broadcasting, a completely unregulated market might fail to allocate resources efficiently. (Setting aside efficiency problems, some would say broadcasting is, in any event, too important in sociocultural terms to be left up to free market—a separate argument.) The most commonly used tools to correct failures have been regulation (e.g., content rules that apply to commercial television and radio operators) and public ownership. Much important research work carried out in the area of media economics over the years has focused on questions around market failure and the merits or otherwise of public ownership as a solution. Economists have put forward different sorts of evidence and arguments concerning how best to advance "public purposes" associated with broadcasting.

For some, the public good characteristics of broadcasting (nonexcludability and nonexhaustability) suggest it would best be supplied by the public sector at zero price, using public funds (Davies, 1999). And indeed, most countries have some sort of publicly funded and state-owned broadcasting entity to provide public service broadcasting (PSB). But in an era of increased choice and when the technology to allow viewers to make payments directly for whatever services they want is well established, others argue that the use of public funds to finance broadcasting is no longer appropriate (Elstein, 2004).

Aside from public ownership, other ways in which state authorities can encourage the dissemination of particular forms of output include provision of public subsidies directed not at organizations but rather at encouraging production or distribution of whatever sort of content is favored. Special support measures that encourage greater supply and consumption of media content that confers positive externalities are very common and have frequently been the subject of analysis in media economics research work.

Policy interventions designed to support media content generally take two forms. Some interventions are essentially protectionist and help domestic producers by restricting the permitted level of imports of competing nondomestic television or feature film content. Work on international trade in audiovisual content, as well as on the dominance of U.S. suppliers and the efficacy of policy measures to counter this, has been a staple in media economics over many years (Noam & Millonzi, 1993). The affect of tariffs, quotas, and trade disputes in the audiovisual sector have received attention from numerous economists, most notably Acheson and Maule (2001) and Hoskins et al. (1997, 2004).

An alternative approach, rather than imposing tariffs or trade barriers, is to provide grants and subsidies for content producers. European countries such as Germany and France have a long tradition of providing grants to television and film producers to boost indigenous production levels. Work in media economics has helped explain how production grants allow the positive gains to society arising from the availability of, say, indigenously made audiovisual content to be internalized by the production firm, thus correcting the failure of the market system to provide an adequate supply of such content. But the dangers that grants may encourage deviations from profit-maximizing behavior and promote a culture of dependency among local producers are also well covered in media economics texts.

One other very significant concern that arises from the free operation of markets and is frequently a focus in work by media economists is monopolization—the accumulation and potential for abuse of excessive market power by individual media firms and organizations. The prevalence of economies of scale and scope in media, as discussed above, creates an incentive toward expansion and diversification by media firms and, in turn, a natural gravitation toward monopoly and oligopoly market structures. So concentrations of media ownership are a widespread phenomenon, and notwithstanding ongoing technological advances affecting distribution, questions about how policy makers should deal with these have long been of interest to those working in media economics.

A key question is the extent to which media expansion strategies give rise to useful efficiency gains and how much they result in the accumulation of excessive market power within and across media industries. A tricky problem facing policy makers is that sometimes expansion and mergers in the media sector will result in both of these outcomes (i.e., expansion makes possible greater efficiency), but at the same time, it facilitates market dominance and therefore poses risks for competition (Doyle, 2002, p. 166). Concerns about the potential for exercise of monopoly power and about suitable measures to accommodate the development of media firms, while enabling competition, remain important themes in policy making that media economics specialists can help shed light on.

## Impact of Technological Change

Because media industries are heavily reliant on technology, a recurring theme for work in media economics is the impact of technological change. From the arrival of the printing press to the era of wireless Internet, processes of media production and distribution have been heavily dependent on and shaped by technology. For media firms, the need to understand, participate in, and capitalize effectively on technological advancements is a constant challenge. For regulators, the task of protecting and promoting public interest concerns associated with mass communication is greatly complicated by ongoing technological change. So a great deal of work in media economics is about, in one way or another, exploring the implications of recent technological advances.

The introduction of digital technology and the growth of electronic infrastructures for delivery of media have been major forces for change in recent years. The spread of digital technology has affected media production, distribution, and audience consumption patterns with knock-on implications for advertising. In the United Kingdom, the Internet accounted for "nearly one in every five pounds of advertising in 2007" (Ofcom, 2008, p. 27).

In broadcasting, digital compression techniques have multiplied the potential number of channels that can be conveyed. Digital technology has allowed for improved and enhanced television and radio services and for a more efficient usage of available spectrum. Much recent work in media economics has examined the implementation of digital broadcasting, looking at, for example, systems of incentives to encourage broadcasters and/or audiences to migrate from analog to digital, as well as at the economic implications and advantages of redeployment of radio spectrum post-digital switchover.

Digitization has facilitated a greater overlap or convergence in the technologies used in broadcasting, telecommunications, and computing and has opened up opportunities for the development of multimedia and interactive products and services. Convergence is encouraging more cross-sectoral activity and conglomerate expansion, raising new questions about the blurring of traditional market boundaries and barriers. Shaver and Shaver (2006, p. 654) observe that, whereas scholars have tended to focus on individual media industries in the past, the challenges posed by cross-media development will require more evolved approaches in future decades.

The Internet is based on digital technology, and its expansion over recent years has had a seismic impact on the whole media industry. Media content is ideally suited to dissemination via this digital infrastructure. But the question of how much the Internet will revolutionize competition in media content provision is debatable. Graham (2001, p. 145) notes that, despite changing technology and widening market access, the economics of content provision and the importance of reputation (or strong brands) favor the predominance of large players. Goodwin (1998) similarly has argued that these fundamental economic characteristics and features that favor the position of large diversified media enterprises remain largely unchanged because of the arrival of digital technology.

However, for large media content suppliers just as much as small ones, the question of how best to take advantage of digital delivery platforms remains problematic, even a decade into the twenty-first century. Newspaper publishers, having followed their readers and advertisers online and adjusted to a more cross-platform approach, have not found it easy to convert their Web-based readership into revenues (Greenslade, 2009). Notwithstanding the growing popularity of online services based on user-generated content, social networking sites, and search engines and the increasing propensity for advertisers to invest in online rather than conventional media, concerns about the economic viability of Internet-based media provision still abound.

Internet-based television opens up the prospect of a further significant widening of market access to broadcast distribution, albeit that poor or unreliable reception and an uncertain legal environment for Internet-based television have to some extent served as deterrents to new market entry (Löbbecke & Falkenberg, 2002, p. 99). Be that as it may, few broadcasters are ignoring the growth of the Internet (Chan-Olmsted & Ha, 2003). As the infrastructure of the Internet continues to improve, research into organizational responses to new delivery technology in the television industry represents another important area for emerging work in media economics.

Some studies have paved the way in exploring how the Internet has affected mass media (Kung, Picard, & Towse, 2008) and problems that media enterprises have faced in establishing viable business models for Internet-based content provision services. However, in a climate of rapid technological evolution, further economic research work is needed to build our understanding of the transformative effect of this interactive delivery platform.

Ever-expanding avenues for distribution and the growth of interactive and cross-platform products and services have increased overall demand for attractive and high-profile content. At the same time, digital technology has reduced audiovisual production costs, opened up more possibilities for user-generated content, and generally made it more economically feasible to produce content aimed at narrow audience segments. But digitization and the growth of the Internet have also introduced new threats. The possibility of widespread intermediation of data (i.e., for reassembling or repackaging content lifted from other Web sites) is a significant threat for online publishers.

Copyright is an important topic affecting the economics of creation and supply of media output. A number of scholars working in the area of cultural economics have provided useful analyses of systems of incentives for authors

to produce creative output that copyright provides. Prominent among these is Ruth Towse (2004), who explains that "ownership of copyrights is likely to be concentrated in enterprises with excessive market power" (p. 60), but on the other hand, potentially high rewards are needed to offset the risks and heavy initial costs involved in supplying creative works.

Digitization and the scope, created by the Internet, for widespread illegal reproduction and dissemination of copyright protected work have made it more difficult for rights owners to capture all the returns due to them. So far, this has affected the music industry more than others, albeit that recent declines in the revenues earned by record companies are attributable to factors other than piracy. Nonetheless, audiovisual material and indeed any information goods that can be conveyed in a digital format are also now increasingly fallible to large-scale electronic piracy. This poses significant potential challenges, for example, to film distributors. For all content creators and rights owners, electronic piracy or illegal reproduction of copyright protected works is now a major concern. Thus, identifying sustainable revenue models for the future represents a major challenge for many media suppliers, as indeed for economists working in this area.

Conducting scholarly work in the area of media economics can be problematic, not least because, on account of the industry's reliance on technology, it is a sector that is almost always in a state of flux. Rather than deterring interest, however, such challenges are a source of attraction that continues to inspire innovative and exciting research work where the tools of economics are deployed in analyzing media issues and problems.

## References and Further Readings


Acheson, K., & Maule, C. (2001). *Much ado about culture.* Ann Arbor: University of Michigan Press.

Albarran, A. (1996). *Media economics: Understanding markets, industries and concepts.* Ames: Iowa State University Press.

Albarran, A., & Dimmick, J. (1996). Concentrations and economies of multiformity in communication industries. *Journal of Media Economics,* 9(4), 41–50.

Alexander, A., Owers, J., & Carveth, R. (Eds.). (1993). *Media economics: Theory and practice.* Mahwah, NJ: Lawrence Erlbaum.

Barwise, P., & Ehrenberg, A. (1989). *Television and its audience.* London: Sage.

Blumler, J., & Nossitor, T. (Eds.). (1991). *Broadcasting finance in transition.* Oxford, UK: Oxford University Press.

Brown, A. (2003). Different paths: A comparison of the introduction of digital terrestrial television in Australia and Finland. *International Journal on Media Management, 4,* 277–286.

Bruck, P. (1993). *Okonomie und Zukunft der Printmedien* [Economics and the future of print media]. Munich: Fischer.

Caves, R. (2000). *Creative industries: Contracts between art and commerce.* Cambridge, MA: Harvard University Press.

Chan-Olmsted, S., & Ha, L. (2003). Internet business models for broadcasters: How television stations perceive and integrate the Internet. *Journal of Broadcasting and Electronic Media, 47,* 597–617.

Chiplin, B., & Sturgess, B. (1981). *Economics of advertising.* London: Advertising Association.

Collins, R., Garnham, N., & Locksley, G. (1988). *The economics of television: The UK case.* London: Sage.

Davies, G. (Chairman). (1999). *The future funding of the BBC: Report of the Independent Review Panel.* London: DCMS.

De Vany, A. (2004). *Hollywood economics: How extreme uncertainty shapes the film industry.* London: Routledge.

Doyle, G. (2002). *Understanding media economics.* London: Sage.

Doyle, G. (2006). Managing global expansion of media products and brands: A case study of FHM. *International Journal on Media Management, 8*(3), 105–115.

Doyle, G., & Frith, S. (2006). Methodological approaches in media economics and media management research. In A. Albarran, S. Chan-Olmsted, & M. Wirth (Eds.), *Handbook of media management and economics* (pp. 553–572). Mahwah, NJ: Lawrence Erlbaum.

Elstein, D. (2004). *Building public value: The BBC's new philosophy.* London: Institute of Economics Affairs.

Florida, R. (2002). *The rise of the creative class: And how it's transforming work, leisure, community and everyday life.* New York: Basic Books.

Fu, W. (2003). Applying the structure-conduct-performance framework in the media industry analysis. *International Journal on Media Management, 5,* 275–284.

Goff, D. (2006). Global media management and economics. In A. Albarran, S. Chan-Olmsted, & M. Wirth (Eds.), *Handbook of media management and economics* (pp. 675–689). Mahwah, NJ: Lawrence Erlbaum.

Gomery, D. (1993). The centrality of media economics. *Journal of Communication, 43,* 190–198.

Goodwin, P. (1998). Concentration: Does the digital revolution change the basic rules of media economics? In R. Picard (Ed.), *Evolving media markets: Effects of economic and policy changes* (pp. 173–190). Turku, Finland: Business Research and Development Centre.

Graham, A. (2001). The assessment: Economics of the Internet. *Oxford Review of Economic Policy, 17,* 145–158.

Graham, A., & Davies, G. (1997). *Broadcasting, society and policy in the multimedia age.* London: John Libbey.

Greenslade, R. (2009, January 5). Online is the future–and the future is now. *Guardian,* p. 4.

Helgesen, J. (2002). The internationalisation of Norwegian newspaper companies. In R. Picard (Ed.), *Media firms: Structures, operations and performance* (pp. 123–138). Mahwah, NJ: Lawrence Erlbaum.

Hoskins, C., McFadyen, S., & Finn, A. (1997). *Global television and film: An introduction to the economics of the business.* Oxford, UK: Clarendon.

Hoskins, C., McFadyen, S., & Finn, A. (2004). *Media economics: Applying economics to new and traditional media.* Thousand Oaks, CA: Sage.

Hutton, W., O'Keefe, A., Schneider, P., Andari, R., & Bakhshi, H. (2007). *Staying ahead: The economic performance of the UK's creative industries.* London: The Work Foundation.

Kung, L. (2008). *Strategic management in the media.* London: Sage.

Kung, L., Picard, R., & Towse, R. (2008). *The Internet and the mass media.* London: Sage.

Lacy, S., & Bauer, J. (2006). Future directions for media economics research. In A. Albarran, S. Chan-Olmsted, & M. Wirth (Eds.), *Handbook of media management and economics* (pp. 655–673). Mahwah, NJ: Lawrence Erlbaum.

Lee, B., & Bae, H.-S. (2004). The effect of screen quotas on the self-sufficiency ratio in recent domestic film markets. *Journal of Media Economics, 17,* 163–176.

Löbbecke, C., & Falkenberg, M. (2002). A framework for assessing market entry opportunities for Internet-based TV. *International Journal on Media Management*, *4*(2), 95–104.

López, J. T. (1985). *Economia de la communicación de masas* [Economics of mass communications]. Madrid, Spain: Grupo Zero.

Murray, S. (2005). Brand loyalties: Rethinking content within global corporate media. *Media, Culture and Society, 27,* 415–435.

Napoli, P. (2003). *Audience economics: Media institutions and the audience marketplace.* New York: Columbia University Press.

Noam, E., & Millonzi, J. (Eds.). (1993). *The international market in film and television programs.* Norwood, NJ: Ablex.

Ofcom. (2008). *The international communication market 2008.* London: Author.

Owen, B., & Wildman, S. (1992). *Video economics.* Cambridge, MA: Harvard University Press.

Peacock, A. (Chair). (1986). *Report of the committee on financing the BBC.* London: HMSO, Cmnd. 9824.

Peacock, A. (1989). Introduction. In G. Hughes & D. Vines (Eds.), *Deregulation and the future of commercial television* (David Hume Institute Paper No. 12, pp. 1–8). Aberdeen, UK: Aberdeen University Press.

Pennington, A., & Parker, R. (2008, November 19). Multiplatform commissioning. *Broadcast,* p. 11.

Picard, R. (1989). *Media economics: Concepts and issues.* Newbury Park, CA: Sage.

Picard, R. (2002). *The economics and financing of media companies.* New York: Fordham University Press.

Picard, R. (2006). Comparative aspects of media economics and its development in Europe and the USA. In H. Jurgen & G. Kopper (Eds.), *Media economics in Europe* (pp. 15–23). Berlin: VISTAS Berlag.

Porter, M. (1998). *Competitive strategy: Techniques for analyzing industries and competitors.* New York: Free Press.

Sánchez-Tabernero, A., & Carvajal, M. (2002). *Media concentrations in the European market, new trends and challenges, media markets monograph.* Pamplona, Spain: Servicio de Publiciones de la Universidad de Navarra.

Schmalensee, R. (1972). *The economics of advertising.* Amsterdam: North-Holland.

Shaver, D., & Shaver, M. A. (2006). Directions for media management research in the 21st century. In A. Albarran, S. Chan-Olmsted, & M. Wirth (Eds.), *Handbook of media management and economics* (pp. 639–654). Mahwah, NJ: Lawrence Erlbaum.

Steemers, J. (2004). *Selling television: British television in the global marketplace.* London: BFI.

Toussaint-Desmoulins, N. (1978). *L'economie de medias* [The media economy]. Paris: Presses Universitaires de France.

Towse, R. (2004). Copyright and economics. In S. Frith & L. Marshall (Eds.), *Music and copyright* (pp. 54–69). Edinburgh, UK: Edinburgh University Press.

Van der Wurff, R., Bakker, P., & Picard, R. (2008). Economic growth and advertising expenditures in different media in different countries. *Journal of Media Economics, 21,* 28–52.

Waterman, D. (2003). Economic explanations of American trade dominance: Contest or contribution? *Journal of Media Economics and Culture, 1*(1).

Webster, J. (2005). Beneath the veneer of fragmentation: Television audience polarization in a multichannel world. *Journal of Communication, 15,* 366–382.

Webster, J., Phelan, P., & Lichty, L. (2000). *Ratings analysis: The theory and practice of audience research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Wildman, S. (2003). Modeling the advertising revenue potential of media audiences: An underdeveloped side of media economics. *Journal of Media Economics and Culture, 1*(2), 7–37.

Wirth, M., & Bloch, H. (1995). Industrial organization theory and media industry analysis. *Journal of Media Economics, 8*(2), 15–26.

# 82

# Microfinance

Shannon Mudd
*Ursinus College*

Over the past several decades, microfinance, broadly defined as financial services to poor and low-income clients, has become an increasingly important tool for governments, multilateral agencies, and nongovernmental organizations (NGOs) to address poverty. Initially, for example, with Banco Sol in Bolivia, the Grameen Bank of Bangladesh, and Bank Rakyat of Indonesia, microfinance was focused primarily on microcredit, small loans to poor people. The basic idea was to extend credit to poor people who do not have access to finance, enabling them to help themselves. In designing products for the poor, the industry has made substantial innovations in the practices used in lending. In addition, some microfinance institutions (MFIs) now offer a range of financial services, including savings vehicles, money transfers, and insurance specifically designed to meet both the needs and specific situations of poor people. Broad recognition of microfinance as a development strategy came with the United Nations (UN) declaring 2005 The International Year of Microcredit and with the awarding of the 2006 Nobel Peace Prize jointly to Mohamed Yunus and to the Grameen Bank, which he founded.

In this chapter, we examine some of the economic questions associated with microfinance, particularly credit.[1] At the most basic level is the question of why the poor have not had access to finance in the past. A surprising outcome of the "microfinance revolution," as it was referred to by Marguerite Robinson (2001), is evidence that poor people, despite their impoverished situation, are good credit risks. Poor people borrowing small amounts of money almost always repay their loans, including sometimes fairly steep interest charges, and do it on time. This suggests that they find productive uses for the funds ("Economics Focus," 2009). But if they are good credit risks, why haven't banks been operating in this sector in the past—that is, what is the market failure? To answer this, we look at how banks function as a response to problems of asymmetric information. Over the past several centuries, banks have developed a number of common practices to address these problems, such as the use of collateral, restrictive covenants in binding loan contracts, credit registries, and so on. However, many of them are not applicable to poor people. The innovative practices developed by microfinance institutions serve as alternative, innovative responses to this same problem.

## Some Brief General Statistics

Before looking directly at the economic questions of microfinance, it is helpful to look at some statistics from the industry. There has been substantial and sustained growth in the microfinance industry. Between 1997 and 2005, the number of microfinance institutions increased from 618 to 3,133. The number of borrowers increased from 13.5 million to 113.3 million, with 84% of them being women (Daley-Harris, 2006). This works out to an average annual increase of nearly 20%. Gonzalez and Rosenberg (2009) find a lower value when adjustments are made to data sources that incorporate institutions for the first time as if their number of borrowers are all new when in fact the institution has had years of history and borrowers. Still, their adjusted figure of an average growth rate of 12% is substantial, although they do note a recent slowing.

The Gonzalez and Rosenberg (2009) data provide some other interesting results. For example, while microfinance is often associated with NGOs, they show that in 2004, these institutions accounted for only 24% of borrowers

(financial cooperatives are not included as there was not sufficient representation in the sample to analyze). Licensed private banks and finance companies accounted for a further 17%, while state-owned institutions accounted for 30%. The remaining 29% was accounted for by the large number of borrowers in Indian self-help groups. As these Indian self-help groups are mostly financed through state banks, adding them to share from state-owned institutions indicates that government-financed organizations account for well over 50% of all borrowers. Note, however, that these statistics exclude financial cooperatives for which there was not sufficient representation in the sample to analyze.

Survey data from Lapenu and Zeller (2001) indicate that only a small percentage of microfinance institutions use the group lending methodology that has been so closely associated with microlending, while most microfinance institutions use individual-based lending. Yet, the 16% of institutions using group lending account for more than two thirds of the actual microfinance borrowers.

Gonzalez and Rosenberg (2009) also report on the profitability, concentration, and geographic distribution of microfinance. Measured by number of borrowers, South Asia dominates with 67 million of the 94 million borrowers in the database. East Asia and the Pacific comprise another 21 million. Sub-Saharan Africa and Latin America comprise 6 and 5 million, respectively. Because of a relatively late start, the regions of the Middle East/North Africa and Eastern Europe/Central Asia have relatively little microfinance. Because advanced economies have wealthier populations and well-developed financial systems, borrowers are a very small percentage of their population.

South Asia is very populous, but even on a per capita basis, it has twice as much microcredit as any other region with nearly 2.25% of the population having microfinance loans. The regions of East Asia and the Pacific and of Latin America have just over 1%, with sub-Saharan Africa slightly lagging.

One of the biggest debates in microfinance in recent years has been the issue of sustainability, the ability of a microfinance organization to maintain its operations without donor funds or subsidization. While most MFIs are not profitable, 44% of borrowers work with profitable MFIs. Interestingly, they find that MFIs tend to be more profitable than the commercial banks operating in the same country, although the data set does not include many tiny MFIs, which would tend not to be profitable. Not surprisingly, profitable MFIs in the database grow faster than those that are not profitable.

Like many industries, especially in developing countries, microfinance tends to be highly concentrated. Within a country, the median share of the largest MFI is one third of the entire market. The median share of the top five microfinance institutions in a country is 81%, and for the top 10 MFIs, it is 95%. This high level of concentration also extends to the world market. Nine percent of the MFIs account for 75% of all microfinance borrowers.

## Microfinance as a Response to Information Asymmetries

> *"Only those with money are able to borrow it." —Proverb*

### Asymmetric Information in the Banking Sector

Like many a proverb, there is some wisdom within. To understand why this proverb rings true as a description of the way finance often works, consider a basic problem of finance, *information asymmetry.* Information asymmetry arises when the agent on one side of a transaction has more information than the agent on the other side *and* the agent with more information cannot easily and credibly convey that information to the other party even if he or she tries.

While standard examples of markets with asymmetric information include the used car market and markets for insurance, we also find the problem of asymmetric information in finance. Consider one of the basic functions of a financial sector, *financial intermediation.* Financial intermediation is the process in which an organization gathers funds from those who do not have immediate productive use for it and channels these funds to those who can use it productively. Banks, the most important financial intermediary, take in deposits and then use these funds to make loans to individuals, businesses, and governments. Traditionally, the banks earn income from the interest rate spread (i.e., the difference between the interest paid to their depositors and the interest earned on loans). Like any firm, they seek to ensure revenues cover their costs. Their costs include not only the interest paid on deposits but also the costs of screening, monitoring, and enforcing loan agreements with borrowers; the costs of providing additional services to clients; overhead expenses; and so on. Banks also incur losses from unpaid loans. To deal with the risk of unpaid loans, the interest rates charged to borrowers reflect the level of *credit risk* (i.e., the risk that the borrower will not pay back the loan).

Consider the two types of credit risk associated with business loans. First, credit risk arises because there is uncertainty about the income the borrower is able to generate from its activities. The borrower may not pay back the loan because of a bad outcome of those activities; for example, demand for a redesigned product may not be as large as expected. The higher the probability of a bad outcome, the higher the probability the borrower will be unable to make its payments and default and the lower the expected income from the loan. Banks charge a higher interest rate to reflect this credit risk.

Second, credit risk stems from asymmetric information because the borrower's actions are not completely

observable. The borrower always knows more about his or her activities than the bank and may engage in behaviors that would negatively affect the bank. For example, suppose Joe the plumber receives a bank loan to purchase a new piece of plumbing equipment that will enable Joe to expand his business and earn greater income. The first type of behavior of concern for the bank is that Joe may simply break the promise to pay back the loan. Another behavior of concern would be if Joe decided not to work as hard once he received the loan. Furthermore, Joe may engage in more risky activities than the bank would desire, for example, by not using the money to purchase new plumbing equipment but instead purchasing a share in a racehorse that his neighbors own. These examples are of behaviors that occur *after* the loan has been made and are referred to as problems of *moral hazard.* Once the loan is made, the incentives for the borrower such as Joe to do what he said he would do may no longer be as strong. If the bank cannot monitor the borrower's behavior, the borrower may, for example, prefer to work less hard since it is now the bank's money at risk, and he can try to renegotiate the payment schedule of the loan claiming he is unable to pay because of a bad economy beyond his control rather than because he did not work hard. If the problem of moral hazard is great, banks may choose not to lend, and there is market failure from this information asymmetry.

The problem of information asymmetry *before* the loan occurs can result in *adverse selection* or the "lemons problem" (Akerlof, 1970). A potential borrower knows more about his or her intentions of keeping a promise than the lender. Because the lender does not know whether a given borrower represents a high risk of breaking its promise, the bank will charge an interest rate that represents the expected level of risk determined by the bank's assessment of the proportion of high-risk and low-risk applicants for loans. However, those who know they are low-risk borrowers may not be willing to pay such a high interest rate and may drop out of the market. As the low-risk applicants drop out, the probability of a loan applicant being high risk increases, the expected risk is higher, and the bank will charge a higher interest rate. This causes even more low-risk applicants to drop out of the market. Thus, there is "adverse selection": The low-risk applicants select themselves out of the market, and only high-risk applicants remain.

As banks generally prefer not to lend to high-risk borrowers, at the extreme, they may stop lending. In this case, the market fails (i.e., there is no lending) because low-risk borrowers cannot credibly identify themselves in this situation of asymmetric information. In a less extreme outcome, banks may choose not to charge high interest rates and instead ration credit to try to avoid asymmetric information problems (Stiglitz & Weiss, 1981). Although it is not complete market failure, it still presents problems for access to finance, especially to the poor. If banks are rationing credit, one must ask what determines who receives the limited amount of loans.

## Standard Banking Responses to Problems of Asymmetric Information

Over time, banks have developed a number of mechanisms of screening, monitoring, and enforcement to try to mitigate problems of adverse selection and moral hazard. These mechanisms are explored here while the following section describes the alternative mechanisms microfinance has developed specifically for the poor.

To deal with the adverse selection problem, banks become experts both in analyzing information produced by the firm and in the further production of information about firms. This enables them to better assess risk and screen out good risks from bad risks. Established firms seeking a loan provide formal accounting data for the bank to assess. However, banks may gather additional information on the background of potential borrowers to determine, for example, if they had been involved in criminal activities in the past. Banks may also seek to collect information on the credit history of potential borrowers to find out whether they have failed to fulfill past promises or been unable to meet payment obligations such as rent, utility bills, credit card payments, and so on. Banks may also gather information about previous business activities to assess the business acumen of the potential borrower. They may also look at the market in which the potential borrower is operating to better assess the borrower's business plans and projections of future costs and sales.

Banks also often require *collateral,* a valuable asset that is surrendered to the bank if the borrower is not able to fulfill its promise to repay the loan (remember the proverb!). Collateral requirements reduce adverse selection as high-risk borrowers who know they are unlikely to pay off the loan are less willing to risk giving up a valuable asset and may choose not to apply. Furthermore, collateral helps to mitigate the costs of adverse selection to the bank when a high-risk borrower defaults. However, even if the bank were to take possession of the collateral, there are likely to be significant costs involved for the bank, and it may not be able to recover the full value of the loan.

To further reduce the other problems of asymmetric information, moral hazard problems, banks engage in monitoring and write debt contracts with restrictive covenants that limit the behavior of borrowers. For example, banks may monitor borrowers by requiring them to regularly report sales volumes, maintain bank accounts at the lending bank, and so forth. They may have a contract that limits how the funds can be used, for example, to purchase a certain type of equipment. These contracts are enforceable through the courts and provide the bank with recourse should the borrower try to use the money for other types of activities that the bank would deem undesirable. Bank loan officers may choose to visit the borrower

to check that the funds are being used as agreed. Of course, collateral also serves to reduce the problem of moral hazard. The loss of the valuable asset continues to provide an incentive to borrowers to fulfill their promise to repay.

Note that these responses to the problems of asymmetric information are standard in "arm's-length" banking systems. An alternative response is "relationship banking," which is more common in Asian countries. In these banking systems, banks develop strong ties to groups of firms, for example, through cross-ownership. Because of the close ties, banks have more intimate knowledge of borrowers and often some control, lessening the problems of information asymmetry. Both systems have advantages and disadvantages. For example, arm's-length banking systems do not work well in countries that do not have strong judicial systems to enforce contracts. The relationship banking system may lead to problems of transparency, due to a lack of information production or because the close ties can lead to cronyism and noneconomic decision making. For further discussion, see Rajan and Zingales (1998).

## The Microfinance Innovation

While, generally, the poor are considered bad credit risks because they lack net worth that might be used to pay off a loan in the event of a bad outcome of a business activity, perhaps more important is the problem that many of the standard responses to the problems of asymmetric information mentioned above are not appropriate for microloans to the poor, particularly in developing countries. First, the poor lack easily valued, marketable assets that could serve as collateral. Furthermore, especially in developing countries, the poor are less likely to engage in activities that would provide relatively accessible background information about them (credit cards, utility payments, etc.). Their business activities may be more informal and lack financial records or business plans. Many developing countries may also not have a well-functioning legal system to enforce contracts. For example, according to the 2008 Doing Business Report, the time required to settle a contract dispute in India is 1,420 days. This is nearly three times the average of Organisation for Economic Co-operation and Development (OECD) countries and represents a significant cost to any party trying to enforce a contractual obligation.

Microfinance has developed a number of innovative practices that serve as alternative responses to asymmetric information problems. These include group lending, dynamic incentives, regular repayment schedules, collateral substitutes, and lending targeted toward women. These are discussed individually below.

## Group Lending

> Once every week in villages throughout Bangladesh, groups of forty villagers meet together for half an hour or so, joined by a loan officer from a microfinance organization. The loan officer sits in the front of the group (the "center") and begins his business. The large group of villagers is subdivided into eight five-person groups, each with its own chairperson, and the eight chairs, in turn, hand over their group's passbooks to the chairperson of the center, who then passes the books to the loan officer. The loan officer duly records the individual transactions in his ledger, noting weekly installments on loans outstanding, savings deposits, and fees. Quick arithmetic on a calculator ensures that the totals add up correctly, and, if they do not, the loan officer sorts out any discrepancies. Before leaving, he may dispense advice and make arrangements for customers to obtain new loans at the branch office. All of this is done in public, making the process more transparent and letting the villagers know who among them is moving forward and who may be running into difficulties.
>
> This scene is repeated over 70,000 times each week in Bangladesh by members and staff of the Grameen Bank, and versions have been adapted around the world by Grameen-style replicators. Other institutions instead base their methods on the "solidarity group" approach of Bolivia's BancoSol or the "village bank" approach operated by microlenders in seventy countries through Africa, Latin America, and Asia (including affiliates of FINCA, Pro Mujer, and Freedom from Hunger). For many, this kind of "group lending" has become synonymous with microfinance. (Armendariz de Aghion & Morduch, 2005, p. 85)

In his innovative, seminal venture into microfinance, Mohammed Yunus recognized that the very low income of the potential borrowers he was targeting meant that collateral, the standard tool used in lending in developed banking systems, was not a mechanism he could use to reduce the basic lending problems of adverse selection and moral hazard. Instead, he devised a way to use the social ties among a group of borrowers to help avoid these problems.

In the traditional Grameen model of group lending, loans are administered to groups of five borrowers who form voluntarily. Loans might be used for rice processing, the raising of livestock, traditional craft materials, and so on. The process of lending starts with two members of the group receiving funds. After these two start making regular payments, loans are gradually extended to two additional members and eventually to the fifth member. In the Grameen model, the group meets with their lender weekly, along with seven other groups, so that a total of 40 group members participate. In this way, the program builds a sense of community, or *social capital,* as well as individual self-reliance. However, just as important is a "joint responsibility" rule in forming the group that if any one member of the group of five defaults, all of the members will be blocked from future access to loans from the lender. Some group lending programs are even more restrictive, requiring "joint liability." In this type of program, group members may be required to make payments in the case of nonpayment or default by one of its members.

Group lending works to avoid the problem of asymmetric information by taking advantage of local information in screening, monitoring, and enforcement. Group members often know each other and can monitor each other relatively

easily. The joint responsibility rule ensures that it is in all group members' interest that each member meet his or her obligations. The group may also serve as a localized enforcement mechanism, able to threaten social isolation (or even physical retribution). Together, these provide a powerful antidote to moral hazard problems. To ensure obligations are met, the group may also serve as an informal insurance contract so that if one member has a bad outcome, such as becoming sick, the others may make the sick member's payments until he or she can return to work.

Local knowledge may also help address the problem of adverse selection. Potential borrowers may be able to distinguish who among them are inherently risky borrowers and who are relatively safe borrowers. If banks knew this information, they could charge higher interest rates to the more risky borrowers to reflect the greater risk of default. By allowing groups to form voluntarily, potential borrowers can sort themselves into groups of relatively risky and relatively safe borrowers. Safe borrowers will seek to stick together. Risky borrowers will have no choice but to form groups with other risky borrowers. Because risky borrowers will have more instances of default, the joint liability rules ensure that they make more payments to cover other members of their group when their risky ventures do not succeed. Members thus effectively pay a higher interest rate than the safe borrowers in their separate groups. This sorting helps transfer the risk from the banks to the risky borrowers. And this occurs without the bank itself uncovering the information and without different contracts for different groups (Armendariz de Aghion & Morduch, 2005; Ghatak, 1999; Morduch, 1999).

However, there may be some drawbacks to group lending when social ties are too strong between friends or relatives, worsening repayment rates. Higher levels of social cohesion may lead to collusion among borrowers to cheat the microfinance lender. In addition, group lending may not work as well for larger loan amounts.

## Dynamic Incentives

A common practice of microfinance institutions is dynamic incentives in extending an initial loan to a borrower for only a small amount but increasing the loan amount over time with successful repayment (Besley, 1995). This is sometimes referred to as "step lending" or "progressive lending." For the borrower, the increasing access to funding provides an incentive to continue to meet obligations and so reduces moral hazard. Furthermore, the repeated aspect of these transactions establishes a long-term relationship between the borrower and the lender, facilitating the MFI's gathering of "soft information" about the borrower.

The effectiveness of dynamic incentives is limited when borrowers are mobile and when there are competing lenders in the area. If a borrower changes location, the advantage of the established relationship is lost, and the borrower no longer has an incentive to pay back the loan to gain a larger subsequent loan. Thus, the value of such dynamic lending may be low in more urban areas in which mobility is high and other lenders are available in neighboring areas. Borrowers who default may find it relatively easy to move to another area and start a new relationship with a lender who is unaware of the previous default. In some areas, MFIs are working to share information to reduce this problem.

## Regular Repayment Schedules

In developed banking systems, the common practice in small business lending is for payment in full (a lump-sum payment) of principal and interest at the end of the loan period. In contrast, microfinance institutions often issue loans with frequent payment schedules that begin soon after the loan is disbursed. For example, the terms for an annual loan may consist of 50 weekly payments of principal and interest that begin 2 weeks after the initial disbursement. Such a regular repayment schedule provides several advantages in dealing with information asymmetries. First, the process serves to screen out undisciplined borrowers who recognize their inability to manage their funds to keep such a schedule of payments. The practice also helps with monitoring of the borrower, by the MFI or by the peer group borrowers, who quickly learn about cash flow issues and can address problems at an early stage.

There may be an additional benefit to the borrower of the commitment mechanism of frequent payments unrelated to asymmetric information. Poor households often have difficulty amassing funds over time due to a number of different types of diversions, including other demands on funds, requests from relations who are aware of the household's availability of funds, and also theft (Rutherford, 2000). This limits the ability of households to collect funds over a longer time to make large payments. The frequent payments force the household to prioritize its funds and help limit the diversions that may have prevented it from amassing the savings, which otherwise would have obviated the need for borrowing.

However, the practice of frequent payments has some drawbacks. The early payments will tend to create a bias in lending to those households who have additional sources of cash income, especially when loan proceeds are used for investments that do not generate immediate cash flows. This problem is particularly apparent in agricultural lending for fertilizer and seeds for crops that will not produce cash flow until a future harvest. As Morduch (1999) points out, the MFI is effectively lending against the household's steady diversified income stream rather than on the project itself and its particular riskiness. This will limit the usefulness of this mechanism for certain populations. (Some MFIs have loan products specifically designed for agriculture and other seasonal industries, such as tourism. The repayment schedule is adjusted to take into account the repayment capacity of the borrower, e.g., repayment may be due only at harvest.)

### Collateral Substitutes

Microfinance loan terms may include collateral or partial collateral in a number of different forms. For example, Grameen Bank required all borrowers to contribute to an "emergency fund" in an amount proportional to the loans received (0.5% beyond a set minimum). This fund serves as insurance for group members against death, disability, and default, with payouts related to the length of membership. In addition, loan disbursements were subjected to a 5% group tax paid into a group fund account. Up to half of this fund could be accessed by members, with unanimous group consent, as a zero-interest loan. While initially not allowed, Grameen Bank now allows these funds to be withdrawn by members who are leaving the group. Although the term *collateral* may not be used, these funds function as partial collateral, and the group fund serves as a form of forced savings.

Some microfinance institutions explicitly require collateral—for example, Bank Rakyat Indonesia's *unit desa* program. While some may point to the fact that collateral is rarely collected to show its relative unimportance, noncollection of collateral does not mean the threat of its collection did not have a big effect on the borrower's repayment behavior and that it does not serve as an important enforcement mechanism. However, it is difficult to parse the influence of collateral from the influence of other practices of a microfinance institution, such as dynamic lending.

### Lending Targeted Toward Women

One interesting practice of many microfinance institutions is to focus on lending to women. While for some MFIs, this may be connected to gender issue goals, such as increased empowerment for women, for many the simple fact that repayment rates for women tend to be higher goes far to explain this focus. However, there is substantial debate about why women are more likely to repay loans than men. Some explanations are quite simplistic, stating that women are simply more reliable and less likely to use funds for nonessential leisure purposes, including tobacco and alcohol. Looking deeper into gender differences, some assert that women are more sensitive to negative reactions of fellow members and loan officers when payment difficulties arise. Another explanation that may be more consistent with information problems is that women are simply more likely to be close to the home and thus more easily located and more easily pressured. Men may work away from the home and may more easily remove themselves from difficult situations, making it more difficult to monitor them.

Other explanations of the difference in repayments behavior hinge on societal differences. For example, in many societies, men, but not women, have alternative sources of credit, whether formal or informal. This access to additional credit beyond the MFI means men are less affected by the removal of access to credit by the MFI. For them dynamic incentives (i.e., step lending) will have less of an impact. A further societal difference is that, in some societies, women may be more involved with the type of small trade that can most effectively use and service microfinance loans.

### A Potential Concern About Competition

The innovations reported above can all be viewed as alternative responses to problems of asymmetric information from the practices employed by the commercial banking industry. There is another information problem that may arise as the industry grows.

As competition in microlending becomes stronger, there is a risk that competitors may poach borrowers. The issuance of a loan from one microlender can serve as a signal to another lender that a borrower is a good credit risk. This can serve as an inexpensive screening mechanism of borrowers. Like the problems of information asymmetry, this information problem can also lead to a market failure. The poaching of clients from other MFIs may reduce the incentive for MFIs with sustainability goals to incur the initial costs of screening a new borrower, meaning competition may lead to reduced lending (Peterson & Rajan, 1995).

## Empirical Findings on Alternative Practices Used in Microfinance

Microfinance has developed a number of innovative lending practices to deal with problems of asymmetric information and the particular situations of the poor. But given there are both drawbacks as well as benefits to the different innovations, which types are most effective? Within each type of innovation, which aspects of the practices are most important? And what is it that makes them work? These are some of the economic questions that arise in examining microfinance innovations.

To answer these questions, ideally, researchers would set up an experiment with a large group of households receiving microfinance using different types of contracts and compare the outcomes. This is difficult to do in the field, especially since MFIs tend to specialize in the types of lending terms they use. While poverty outcomes are certainly the primary interest, repayment rates have been the primary focus of empirical research, with many of the studies looking specifically at the effectiveness of group lending.

Some researchers have conducted experiments in a lab setting to better understand how group lending works (Abbink, Irlenbusch, & Renner, 2006). A key question is the importance of social ties on repayment rates. The experiment compared outcomes when groups were formed randomly to outcomes when groups were self-selected, presuming that the latter were largely among people with

preexisting social ties. The results indicate that the groups with stronger social ties performed no better (and sometimes worse) in terms of repayment rates than groups with no social ties.

This result seems counter to the standard description of group lending that relies on social ties for selection, monitoring, and imposing social sanctions as an enforcement mechanism. However, field studies of group lending in Guatemala (Wydick, 1999) and in Thailand (Ahlin & Townsend, 2003) are consistent with the experimental results, finding that strong social ties have little or even a negative impact on repayment rates. In contrast, studies of "social capital" in Peru (Karlan, 2007) and "social cohesion" in Costa Rica (Wenner, 1995) find positive effects of their increase. It appears that the definitions of social ties used may be key to the findings. While Wydick (1999) found a negative impact of increased social ties (friendship), his measure of social cohesion, proxied specifically by living proximity or knowing each other prior to joining the group, had a positive impact. This is consistent with the earlier mentioned problem that social ties that are too close may be counterproductive.

Note that even with what seem to be good data, evaluating the effects of various microfinance practices is difficult because of selection and other problems. Armendariz de Aghion and Morduch (2005) specifically point out the difficulties of assessing group lending by discussing in more detail the Gómez and Santor (2003) study of two Canadian microlenders that make loans both individually and in groups. Because most microfinance is targeted to the poor in developing countries, to some extent, these programs differ from the norm. With much more developed financial systems, legal systems, and so on, microfinance in more developed countries operates in a very different paradigm, as can be seen by default rates much higher than usually reported. Still, both the methodology and the findings are instructive.

In the two programs studied, while interest rates and fees are similar, there are substantial differences in the populations and loan sizes between the portfolio of individual-based loans and the portfolio of group lending loans. Individual loans tend to be larger ($2,700 vs. $1,000). Group borrowers tend to be female, Hispanic, and immigrant while individual borrowers tend to be male, Canada born, and of African descent. A comparison of the two portfolios shows that group lending loans are more likely to be repaid with 20% default rates versus 40% default rates for the individual loans. However, before attributing the differences in repayment rates to the loan methodology, it is important to take into account the differences in the populations.

> The approach taken by Gómez and Santor is to follow the "matching method" approach of Rosenbaum and Rubin (1983). Using a sample of almost 1,400 borrowers, the method involves first pooling all of the data and estimating the likelihood that a borrower will have a group loan (rather than a standard individual loan). Determinants include age, income, neighborhood, education level, and ethnicity. The estimates yield an index of the probability of taking a group loan, with the important feature that borrowers within the same level of the index also have similar observed characteristics. Reliable comparisons are thus achieved by comparing only borrowers with similar levels of the index. . . . Using this method, Gomez and Santor find that borrowers under group contracts repay more often. The result, they argue, arises both because more reliable borrowers are more likely to choose group contracts and because, once in the group contracts, the borrowers work harder. (Armendariz de Aghion & Morduch, 2005, pp. 104–105)

However, as Armendariz de Aghion and Morduch (2005) point out, this methodology only works if the variables used to develop the index are the only variables that matter in the choice of contract. If there are important omitted variables, the method will not produce consistent results. For example, suppose that borrowers who are higher risk tend to have a relative preference for individual loan contracts as opposed to group lending programs. If so, more of the high-risk borrowers would end up borrowing with individual loan contracts. If these high-risk borrowers are more likely to default, then individual loan contracts will have higher default rates. However, this would not be because of the contract design but because of the selection bias in the type of borrower who chooses the individual loan contract. In this case, if unable to identify high-risk from low-risk borrowers, an interpretation of results from an econometric model that indicates group lending increases repayment rates would be spurious.

Interestingly, further results from the experiments of Abbink et al. (2006) find that groups had higher repayment rates than would be expected if loans were to individuals. They also find that larger groups do worse.

As to the effectiveness of other practices, results are fewer. However, Abbink et al. (2006) do find that women are more reliable.

## Evaluating the Impact and Effectiveness of Microfinance

The experiments and field studies just detailed shed some light on the effectiveness of group lending and other practices on MFI performance, particularly repayment rates. However, the ultimate question for microfinance is not about MFI performance but about how the lives of poor people are affected. At the most basic level, economists ask what are the full social costs and benefits of using microfinance as a development strategy targeted toward the poor. In terms of benefits, how has microfinance affected the lives of borrowers and their families? Does the use of microfinance services lead to increased income and standards of living? Perhaps just as important, does it lead to less vulnerability to negative shocks/bad events,

especially those out of their control (weather, job loss, illness, family death)?

Although microfinance institutions provide many examples of borrowers who are doing quite well, these stories are only anecdotal. A more complete assessment is needed as even on the anecdotal side, there are concerns. For example, Montgomery (1996) worries that group lending puts such high pressure to repay on poor households that it may actually make them worse off in the event of a negative outcome, compelling them to pay even when difficulties arise beyond their control. Examples of how households can be harmed include stories of forced seizure of household utensils, livestock, and so forth of defaulting members.

To fully assess the benefits, one must understand the goals of the microfinance sector. Unfortunately, not all participants agree. While the issues of income or standard of living raised above are certainly an important area of concern, some MFIs view their role as more than just providers of financial services. Their goals may extend to providing health education, increasing the education of children, improving health and nutrition, empowering women, basic business training, and so on. Given the multiple goals, assessment of the success of a microfinance program involves more than an examination of repayment rates and requires a variety of data.

### *Measurement Issues*

The number of good, academic studies of the impact of microfinance on poor people is few but growing, and some are discussed below. Importantly, the industry itself is making an effort to improve impact assessment. Hashemi, Foose, and Badawi (2007) present the efforts of the Social Performance Task Force, which met in Paris in 2005 and agreed on five Dimensions of Social Performance to guide them in designing standards for reporting to better analyze the success of microfinance in positively affecting people's lives. The difficulty of such an assessment is apparent in the list itself. The five dimensions are as follows:

1. *Intent and Design.* What is the mission of the institution? Does it have clear social objectives beyond providing access to credit and other financial services to poor people?
2. *Internal Systems and Activities.* What activities will the institution undertake to achieve its mission? Are systems designed and in place to achieve those objectives?
3. *Output.* Does the institution serve poor and very poor people? Are the products designed to meet their needs?
4. *Outcome.* Have clients experienced social and economic improvements?
5. *Impacts.* Can these improvements be attributed to institutional activities?

For a microfinance institution to collect data on all five dimensions is a large task. MFIs have increased their reporting capabilities on the financial side to gain greater access to both donor and commercial funding. However, this is concentrated only in the first three dimensions. The last two dimensions necessary to evaluate their social performance are much more difficult. Data collection is expensive, and an MFI may find it hard to justify expending its limited funds on this activity when the apparent need among its potential borrowers is so great. However, to fully attribute any impacts on households of borrowing and other financial and nonfinancial services, researchers must collect data on other household factors that can affect outcomes as well as from other similar households that did not use microfinance services. Surveys to collect such data require careful design and can be expensive to implement. While the industry's efforts will surely provide more extensive data to analyze questions, it is likely that academic and donor-sponsored surveys will continue to be key to assessment of the industry's effectiveness in combating poverty.

## The Debate on Sustainability

As noted, data collection to assess financial performance has greatly improved, especially as many MFIs have sought access to private capital. Private capital requires a return on its investment, and to attract this type of funding, MFIs must show that their activities are sustainable (i.e., that borrower repayments are sufficient to be able to repay invested funds). However, there is continuing controversy about whether microfinance institutions should have as a goal to cover all their economic opportunity costs.

For example, Richard Rosenberg (2008) discusses an actual debate at the World Microfinance Forum in Geneva in 2008 between Mohamed Yunus and Michael Chu, a former investment banker and president of ACCION, one of the larger microfinance institutions, and now on the faculty at Harvard Business School:

> The debaters argued about whether commercialization (let's define it as the entry of investors whose primary motive is financial rather than social) is good for microfinance. Yunus thinks that it's immoral to make money off the poor, and that the only kinds of investors needed in microfinance are ones who are willing to accept very limited profits for the sake of keeping as much money as possible in the pockets of the borrowers. Michael thinks that we can't meet the worldwide demand for poor people's financial services unless we can draw in private, profit-oriented capital, and that eventual competition can be counted on to bring interest rates and profits down to consumer-friendly levels in most markets.

In terms of social performance, to the extent that MFIs are profitable and self-sustaining, the need for a full assessment may not be necessary. Even those that rely on volunteer donations may not need a full assessment of costs and benefits if the anecdotal stories of how microfinance loans have improved the lives of borrowers are enough to satisfy potential donors. However, much of the

sector continues to exist with the help of grants and government/taxpayer subsidies. Given scant resources from institutional donors and governments, it is important to determine whether the benefits outweigh the costs and whether these funds are well allocated or could produce a better outcome deployed on some other type of poverty reduction program.

### Empirical Findings

The amount of well-designed studies on social performance is few, but an increasing number of good ones are providing interesting results. One important series of studies is based on a survey of 1,800 households in Bangladesh. For example, Pitt and Khandker (1998) find a number of positive impacts of microlending to households. Interestingly, the impacts differ by whether the loan is made to women or men. One important measure to assess is how consumption is affected. Pitt and Khandker find that household consumption increases by a substantial 18 taka for every 100 taka lent to women. However, the increase in household consumption when the loans are made to men is only 11 taka for every 100 taka lent. Part of the difference is due to how labor choices are affected. While lending has no effect on labor supply by women, men choose to increase their consumption of leisure (i.e., work less). Results on the schooling of children indicate that schooling increases for boys no matter the gender of the borrower. However, whether schooling for girls increases when women borrow depends on the lending program, perhaps an indicator that the nonfinancial emphasis of microfinance programs can have an important affect on their impacts. McKernan (2002) finds that nonfinancial aspects of microfinance programs do have an impact. She finds that the provision of the loan itself only accounts for about half of the measured increase in self-employment income from borrowings. She attributes the other half of the increase to improved borrower discipline, empowerment, and even shared information from the social network that arises from the social development programs accompanying the borrowing, such as vocational training and education about health and other issues. Pitt and Khandker (2002) also find that lending helps households to smooth consumption across seasons, indicating that entry into the programs may be motivated by insurance concerns.

However, the results are not unanimous. For example, Morduch (1999) reports on results using a subset of the same data limited to what he considers comparable households. He finds no effects on consumption or education, although his findings do indicate lending helps with consumption smoothing. Importantly, Pitt (1999) has mounted a strong defense of their original methodologies and results.

Fortunately, as noted before, there are substantial efforts to increase the availability of standardized data from microfinance institutions as well as surveys to better distinguish the effects of different types of contracts, nonfinancial program aspects, and so on on poor household outcomes. Future studies using this and other data will be able to provide additional evidence on the impact and effectiveness of microfinance in addressing issues of poverty. The industry is relatively young and still evolving. The impacts of microfinance may not be fully apparent during the relatively short time periods of most studies. As more data become available, the prospects for future research in this area are great.

For additional information on developments in microfinance there are a number of excellent Web sites devoted to the industry. For example, the UN created a Web site in conjunction with its Year of Microcredit (www.yearofmicrocredit.org). The Consultative Group to Assist the Poor (CGAP) maintains both a general Web site (www.cgap.org) and one that is targeted toward the microfinance community (www.microfinancegateway.org). The Microfinance Information Exchange (MIX) provides analysis and data (www.themix.org). The Grameen Foundation has created measures to facilitate microfinance institutions in working with their clients, the Progress Out of Poverty Index (www.progressoutofpoverty.org). A documentary on microfinance produced by PBS, "Small Fortunes," along with other material, is available at www.pbs.org/kbyu/smallfortunes. In addition, two periodicals focused on microfinance are available online, the Asian Development Bank's *Finance for the Poor* and the Microfinance Information Exchange's *The Microbanking Bulletin.*

## Note

1. For a more thorough analysis of the many economic aspects of microfinance and empirical evidence its effectiveness, see the excellent and accessible Armendariz de Aghion and Morduch (2005).

## References and Further Readings

Abbink, K., Irlenbusch, B., & Renner, E. (2006). Group size and social ties in microfinance institutions. *Economic Inquiry, 44,* 614–628.

Ahlin, C., & Townsend, R. (2003). Using repayment data to test across models of joint liability lending. *Economic Journal, 107,* F11–F51.

Akerlof, G. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics, 84,* 488–500.

Armendariz de Aghion, B., & Morduch, J. (2005). *The economics of microfinance.* Cambridge: MIT Press.

Asian Development Bank, *Finance for the Poor,* quarterly newsletter of Microfinance (note selected bibliography in each issue): http://www.adb.org/Documents/Periodicals/Microfinance/default.asp

*Asymmetric Information* discussed in the context of the work of the three 2001 Nobel Prize laureates in economics: http://www.nobel.se/economics/laureates/2001/public.html

Besley, T. (1995). Savings, credit, and insurance. In T. N. Srinivasan & J. Behrman (Eds.), *Handbook of development economics* (Vol. 3A, pp. 2125–2207). Amsterdam: Elsevier.

CGAP Consultative Group to Assist the Poor: http://www.cgap.org/p/site/c

CGAP's Microfinance Gateway: http://www.microfinancegateway.org

Daley-Harris, S. (2006). *State of the Microcredit Summit Campaign report 2006.* Washington, DC: Microcredit Summit Campaign.

Economics focus: A partial marvel. (2009, July 18). *The Economist,* p. 76.

Ghatak, M. (1999). Group lending, local information and peer selection. *Journal of Development Economics, 60,* 27–50.

The Global Development Research Center: http://www.gdrc.org/icm

Gómez, R., & Santor, E. (2003). *Do peer group members outperform individual borrowers? A test of peer group lending using Canadian micro-credit data* (Working Paper 2003–33). Ottawa, Ontario: Bank of Canada.

Gonzalez, A., & Rosenberg, R. (2009). *The state of microfinance: Outreach, profitability, poverty, findings from a database of 2600 microfinance institutions.* Retrieved from http://ssrn.com/abstract=1400253

Hashemi, S., Foose, L., & Badawi, S. (2007). *Beyond good intentions: Measuring the social performance of microfinance institutions* (CGAP Focus Note No. 41). Washington, DC: CGAP.

Hermes, N., & Lensink, R. (2007). The empirics of microfinance. *Economic Journal, 117,* F1–F10.

IFC Doing Business Report (cross-country comparisons of various measures of the ease of doing business, from setting up a firm, to trading across borders, to hiring and firing employees, to obtaining credit): http://www.doingbusiness.com

Karlan, D. (2007). Social connections and group banking. *Economic Journal, 117,* F52–F84.

Lapenu, C., & Zeller, M. (2001). *Distribution, growth and performance of microfinance institutions in Africa, Asia and Latin America* (FCND Discussion Paper 114). Washington, DC: International Food Policy Research Institute (IFPRI).

McKernan, S.-M. (2002). The impact of micro-credit programs on self-employment profits: Do non-credit program aspects matter? *Review of Economics and Statistics, 84,* 93–115.

Microfinance Information Exchange (MIX), *The Microbanking Bulletin,* semi-annual: www.themix.org/microbanking-bulletin/microbanking-bulletin

Microplace.com: http://www.microplace.com

MIX Market—Global Microfinance Information Platform: http://www.mixmarket.org

Montgomery, R. (1996). Disciplining or protecting the poor? Avoiding the social costs of peer pressure in micro-credit schemes. *Journal of International Development, 8,* 289–305.

Morduch, J. (1999). The microfinance promise. *Journal of Economic Literature, 37,* 1569–1614.

PBS Small Fortunes Documentary and Accompanying Web site: http://www.pbs.org/kbyu/smallfortunes

Peterson, M., & Rajan, R. (1995). The effect of credit market competition on lending relationships. *Quarterly Journal of Economics, 110,* 407–443.

Pitt, M. (1999). *Reply to Jonathan Morduch's "Does microfinance really help the poor? New evidence from flagship programs in Bangladesh* (Unpublished mimeo). Retrieved from http://www.pstc.brown.edu/~mp/reply.pdf

Pitt, M., & Khandker, S. (1998). The impact of group-based credit on poor households in Bangladesh: Does the gender of participants matter? *Journal of Political Economy, 106,* 958–996.

Pitt, M., & Khandker, S. (2002). Credit programs for the poor and seasonality in rural Bangladesh. *Journal of Development Studies, 39*(2), 1–24.

Progress Out of Poverty Index, Grameen Foundation: http://www.progressoutofpoverty.org

Rajan, R., & Zingales, L. (1998). Which capitalism: Lessons from the East Asian financial crisis. *Journal of Applied Corporate Finance, 11*(3), 40–48.

Robinson, M. (2001). *The microfinance revolution.* Washington, DC: The World Bank.

Rosenberg, R. (2008, October 14). Muhammad Yunus and Michael Chu debate commercialization [Web log post]. Available at http://www.microfinance.cgap.org

Rutherford, S. (2000). *The poor and their money: An essay about financial services for poor people.* Delhi: Oxford India Paperbacks and the Department for International Development. Retrieved from http://www.uncdf.org/mfdl/readings/PoorMoney.pdf

Stiglitz, J., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *American Economic Review, 71,* 393–410.

UN International Year of Microcredit: http://www.yearofmicrocredit.org

Wenner, M. (1995). Group credit: A means to improve information transfer and loan repayment performance. *Journal of Development Studies, 32,* 263–281.

Wydick, B. (1999). Can social cohesion be harnessed to repair market failures? Evidence from group lending in Guatemala. *The Economic Journal, 109,* 463–475.

# 83

# Latin America's Trade Performance in the New Millennium

Maritza Sotomayor
*Utah Valley University*

Latin America has witnessed impressive economic growth since 2003. The average annual growth of the gross domestic product (GDP) for Latin America has been 5% for the 2003–2007 period, the highest in three decades. Trade performance has been one of the essential factors in explaining this growth. The purpose of this article is to analyze the main elements of Latin America's trade performance from the 1990s until recent years and to explain whether global demand and economic reforms of the 1990s helped strengthen the role of the region's trade sector.

Latin American countries have had both a positive and negative relationship with the global economy. After their independence around the beginning of the nineteenth century, most of them were involved in the first period of globalization. During this first era of globalization, Latin America based its economic growth on trade of basic commodities with Europe (for some countries, exports represented more than 60% of GDP). With World War I and the Great Depression, international prices of commodities fell, affecting the Latin American countries. An inward-looking economic model known as import substitution industrialization (ISI) was pursued to eliminate industry's growing dependence on imports and the negative impacts of external shocks from the global economy. Although ISI was crucial in the construction of a manufacturing industry for countries such as Mexico, Brazil, and Argentina, there were also inconsistencies in industrial and trade policies. In the end, this had consequences for the whole economy. After three decades of ISI, it was clear for Latin American governments that it was not possible to uphold an industry based on a growing demand for imported inputs relying on foreign debt. Throughout the 1970s, Latin America obtained low-interest loans. However, after the oil crisis in 1973 and the economic depression through those years, real interest rates increased. As a consequence, Latin American countries entered into a decade of debt crisis. The period 1980–1989 is known as *the lost decade* because of Latin America's poor economic performance, increase in poverty, and lack of viable solutions.

Latin American countries began a process of trade liberalization after several stabilization programs. The positive numbers in the trade account (exports minus imports) during the first years of the twenty-first century were the result of two factors. The first was the economic reforms whereby exports were again the growth engine during the 1990s. The second factor was the fact that the positive global economic conditions enhanced commodity prices, which made Latin America again dependent on the volatilities of the foreign market.

This second era of globalization set new challenges for Latin America. The current global economic conditions with the financial crisis will very likely slow down the region's economic growth. However, it can be seen as a juncture to reevaluate Latin America's trade specialization in the global economy. This change would imply diversification of export destinations, increase the participation of manufacturing products in exports, and consolidate efforts for regional economic integration.

This chapter consists of three sections: (1) trade liberalization process, which reviews the main aspects of economic reforms in the 1990s; (2) trade performance in Latin America 2000–2007, which includes analyses of statistics referring to exports and imports (since demand of exports

from China has been an important element in trade performance, there is a subsection regarding the trade between Latin America and China); and (3) the financial crisis of 2008 and trade performance, which evaluates how the global economy's slowdown would affect the region.

## Trade Liberalization Process

This section covers the main elements of Latin America's trade liberalization process that started for most countries in the 1990s. It will give a perspective on how economic reforms transformed the trade sector where exports had become the main engine of growth.

### Theory

Regarding the potential benefits of trade liberalization, Krugman and Obstfeld (2000) stated that an economic opening through tariff reduction eliminates distortions for consumers and producers with the consequent increase in national welfare. An increase in competition through the supply of imported goods in the domestic market pushes prices down with direct benefits for consumers. Dornbusch (1992) pointed out that trade liberalization brings potential dynamic benefits, for example, through the introduction of a new method of production, the opening of a new market, or the carrying out of new production methods. With the same approach, Rodrik (1999) affirmed that benefits from opening markets are not on the side of exports but rather on the side of imports. According to him, developing countries can benefit from imports of capital and intermediate inputs that are too expensive to produce locally.

One of the main criticisms of free trade and its application to developing countries has been that models of free trade suppose perfect competition in the market. However, Latin America's industries are imperfect market structures. Therefore, selective tariffs on industries can prevent additional distortions in the economy. This is referred to as the market failure justification for infant industry protection (Krugman & Obstfeld, 2000).

This argument against trade liberalization in Latin America was used to emphasize its potential negative effects on local industry. If the speed of trade liberalization were too fast (i.e., the tariff reduction is not gradual), the industry would not have time to adjust to upcoming competition after a period of high protection. The results would be the loss of jobs and wiping out of small and medium enterprises that probably could not compete with multinational corporations.

According to Ffrench-Davis (2005), the competitiveness of the industrial sector should not be based on low wages, government subsidies, or tax exemptions. Industrial policies should aim to increase productivity in the export sector. The export sector promoted by these policies should generate value added and spillovers to the rest of the economy (activities such as the *maquiladoras,* where the value added and spillovers are small, could not be part of these efforts).

These different points of view about benefits and cost of trade liberalization provide an introduction for Latin America's economic policies during the 1990s.

### Economic Reforms of the 1990s

Most Latin American economies used different stabilization programs as a result of the debt crisis. The decade of the 1980s was known as *the lost decade* because of the severe impact on GDP per capita (the average annual percentage change of real GDP per capita was –0.4% for the region), inflation (Argentina: 385% and Brazil: 228% both in 1985 or Bolivia: 2,252% for the 1980–1985 period), and the increase of poverty (39% of Latin America's population lived in poverty in 1990 compared with 35% in 1980; Helwege, 1995). At the end of the 1980s, most of Latin America engaged in economic reforms that had exports as the main engine of economic development. A second era of globalization began with economic reforms that included privatization (reducing the role of the state in the economy) and trade liberalization (opening the economy).

These economic reforms were part of what is now termed the "Washington Consensus" reforms or the neoliberal policies (free-market approach). Williamson (1990) summarized in an article the 10 policy instruments that institutions such as the U.S. Treasury, the International Monetary Fund (IMF), the World Bank, and the Inter-American Development Bank (all of them located in Washington, D.C.) advised as the most significant economic reforms for Latin America. These were fiscal discipline, reordering public expenditure priorities, tax reform, liberalizing interest rates, a competitive exchange rate, trade liberalization, liberalization of foreign direct investment (FDI), privatization, deregulation, and property rights (Williamson, 1990).

Throughout this period, several Latin American countries entered into a period of democratic processes (e.g., elected presidents Alfonsin in Argentina, Siles in Bolivia, Collor in Brazil, and Belaunde in Peru), which helped with the implementation of these economic policies. Besides reduction of the role of the state, economic policies pursued an increase in international trade's participation and promotion of private investment. Argentina followed closely the Washington Consensus recommendations. However, this country suffered from a peso crisis and a reduction of GDP (as a consequence, unemployment levels reached record levels of 20% in 1995; Hofman, 2000). Countries such as Brazil decided not to follow the neoliberal approach. It established a middle path with changes in a slow pace with different stabilization plans (heterodox reforms).

## Trade Performance in Latin America 2000–2007

### Main Economic and Social Characteristics

Table 83.1 summarizes the main economic and social characteristics of Latin American economies for recent

years. These data offer an outlook on similarities and differences among these nations.

Generally speaking, Brazil and Mexico account for more than half of the total region's population. Latin America's land area is bigger than the size of the United States and Canada together. The land areas range from El Salvador with 8.1 thousand square miles to Brazil with 3.3 million square miles. Natural resources are diverse because of geographical and geological conditions. The Andes cross South America from north to south where mineral resources are relatively abundant in countries such as Chile, Peru, and Bolivia. The tropical climate of Colombia is suitable for coffee production, while the conditions of Argentina's plains are good for the cattle industry.

As can be seen in Table 83.1, Argentina, Brazil, and Mexico together explain three quarters of Latin America's

**Table 83.1** Basic Economic and Social Indicators for Latin America

| | *Population 2007 (in Thousands of Persons)* | *Land Area (in Thousands of km²)* | *Share of Regional GDP 2000–2007 (%)* | *GDP Growth 2003–2007 (%)* | *GDP Per Capita 2007 (in 2000 U.S. Dollars)* | *Urban Population % Total Population 2005* | *HDI 2006 (%)[1]* |
|---|---|---|---|---|---|---|---|
| Argentina | 39,356 | 2,767 | 13.28 | 8.83 | 9,396 | 91.80 | 0.860 (46) |
| Bolivia | 9,828 | 1,099 | 0.42 | 4.13 | 1,090 | 64.23 | 0.723 (111) |
| Brazil | 192,645 | 8,512 | 31.79 | 3.93 | 4,216 | 83.40 | 0.807 (70) |
| Chile | 16,604 | 757 | 3.89 | 4.99 | 6,127 | 86.56 | 0.874 (40) |
| Colombia | 46,116 | 1,139 | 4.85 | 5.90 | 2,843 | 76.62 | 0.787 (80) |
| Costa Rica | 4,475 | 51 | 0.83 | 6.54 | 5,085 | 62.61 | 0.847 (50) |
| Cuba | 11,248 | 111 | n.d. | 8.02 | 4,173 | 76.15 | 0.855 (48) |
| Dominican Republic | 9,749 | 49 | 1.22 | 5.89 | 3,464 | 65.49 | 0.768 (91) |
| Ecuador | 13,601 | 248 | 0.85 | 4.79 | 1,624 | 62.82 | 0.807 (72) |
| El Salvador | 7,108 | 21 | 0.64 | 3.21 | 2,252 | 57.84 | 0.747 (101) |
| Guatemala | 13,344 | 109 | 0.86 | 4.00 | 1,665 | 49.97 | 0.696 (121) |
| Haiti | 9,602 | 28 | 0.16 | 0.83 | 392 | 41.75 | 0.521 (148) |
| Honduras | 7,176 | 112 | 0.38 | 5.88 | 1,420 | 47.83 | 0.714 (117) |
| Mexico | 106,448 | 1,958 | 30.36 | 3.32 | 7,094 | 76.50 | 0.842 (51) |
| Nicaragua | 5,603 | 130 | 0.20 | 3.95 | 885 | 56.99 | 0.699 (120) |
| Panama | 3,337 | 77 | 0.61 | 7.80 | 5,206 | 65.77 | 0.832 (58) |
| Paraguay | 6,120 | 407 | 0.35 | 4.40 | 1,467 | 58.50 | 0.752 (98) |
| Peru | 27,894 | 1,285 | 2.77 | 6.47 | 2,751 | 72.66 | 0.788 (79) |
| Uruguay | 3,332 | 177 | 0.91 | 7.01 | 7,255 | 91.95 | 0.859 (47) |
| Venezuela RB | 27,460 | 912 | 5.65 | 7.92 | 5,789 | 92.77 | 0.826 (61) |
| **Latin America** | **561,046** | **19,949** | **100.00** | **4.94** | **4,723** | **77.84** | **0.810**[2] |

SOURCES: Based on data from Economic Commission for Latin America and the Caribbean (ECLAC), CEPALSTAT Databases and Statistical Publications (http://www.eclac.cl/estadisticas/default.asp?idioma=IN) and Statistical Update (http://hdr.undp.org/en/media/HDI_2008_EN_Tables.pdf).

NOTES: 1. Human Development Index (HDI) was developed by United Nations Development Programme (UNDP) in 1990. According to them, it "is a new way of measuring development by combining indicators of life expectancy, education attainment and income into a composite human development index. The HDI sets a minimum and a maximum for each dimension, called goalposts, and then shows where each country stands in relation to these goalposts expressed as a value between 0 and 1." Numbers in parentheses show the country's ranking. 2. Includes the Caribbean.

GDP, which justifies why these economies share a central role in the region's economic performance. One of the most notable aspects in Latin American economies has been their GDP growth for the 2003–2007 period (with the exception of Haiti with a GDP growth of less than 1%). In Table 83.1, many economies show percentages of growth above 5%, while the percentage for the 1999–2002 period was only 1%. The rapid increase in commodity prices is among the factors that explain this performance, due to an increase in demand by countries such as China and India. Although the region is not homogeneous in per capita income, Argentina and Uruguay have one of the highest levels of income per capita. Haiti is among the lowest income per capita countries in Latin America with severe poverty. Overall, Central America and the Caribbean countries have the worst poverty. This assessment turns out to be more evident with the Human Development Index (HDI) indicator proposed by the United Nations Development Programme (UNDP) in 1990. The index's objective is to add social aspects such as life expectancy, education attainment, and income to the existing economic indicators to give a broader measure of economic development. The HDI goes from 0 to 1, and countries are ranked according to this index (UNDP, 1990). The last column of Table 83.1 includes the HDI and the country's ranking in parentheses for 2006. According to this index, Chile, Argentina, and Uruguay have shown high levels of economic development. Even though Uruguay is a small country in comparison with the rest of the region, this country can be considered as developed in terms of its HDI. Then again, although Brazil has a big economy, its HDI was below average for the overall region. As can be inferred from Table 83.1, Latin America is not a homogeneous region in social and economic terms. The trade liberalization process initiated in the 1990s allowed these countries to participate in the global economy. However, the results from that experience have not always translated into an increase in living standards.

## Tariff Reduction and Regional Trade Agreements

One of the policy tools used in the trade liberalization process was the reduction of tariff rates. Figure 83.1 shows Latin America's simple average tariff for manufacturing products for the 1980–2006 period and for selected countries.

Protectionist measures taken throughout ISI led to an increase in tariff rates that lasted for several decades. In 1980, Brazil had an average tariff for manufactured imports of 99.4%, while the rest of Latin America averaged a rate of 50%. The exception was Chile since this country started opening its economy in the mid-1970s. This opening process implied for some countries joining the General Agreement on Tariffs and Trade (GATT), which in 1995 became the World Trade Organization (WTO). As members, they have to limit their tariff rates. In consequence, for 2006, the tariff rates fluctuated around 10% or less.

During the 1990s, there was a revival effort for regional economic integration as an element of trade liberalization in Latin America. The first experiences with regional trade integration occurred in the 1960s. However, attempts to form regional integration arrangements did not reach the expected results. One of the first integration experiments took place when some Central American countries (Costa Rica, El Salvador, Honduras, Nicaragua, and Guatemala)

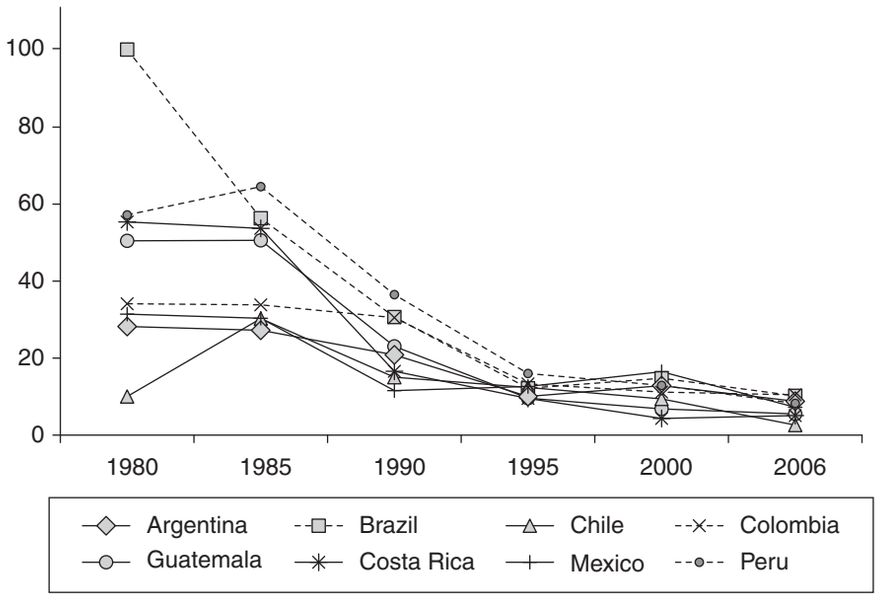


**Figure 83.1** Simple Average Tariff for Manufacturing Products in Latin America 1980–2006 (%)

SOURCES: Based on data from the United Nations Conference on Trade and Development (UNCTAD, 2008); World Trade Organization, *World Tariff Profiles,* various years; Gwartney and Lawson (2008); and data from http://www.freetheworld.com.

together formed the Central American Common Market (CACM) in 1960. It was created as a "custom union." A "custom union" is a free trade area with a common external tariff for nonmembers. According to Bulmer-Thomas (2003), one of CACM's problems was the size of its market and the inward-looking industrial policies that hurt exports of manufacturing products.

The Andean Community treaty was signed in 1969. Its objective was to build a common market between Bolivia, Colombia, Ecuador, Peru, Venezuela, and Chile. The Andean Community has been in place for four decades. It has not reached the stage of a common market. It has faced problems of coordination, and the institutions have not worked properly.

Other experiences with common markets occurred when the Caribbean countries formed the Caribbean Community and Common Market (CARICOM) in 1973, signed by all Caribbean countries. The Latin American Integration Association (LAIA) was formed in 1980 as a free trade area with most South American countries (Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay, and Venezuela), as well as Cuba and Mexico.

The first regional integration agreement of the 1990s was formed with Argentina, Brazil, Paraguay, and Uruguay as members of the Southern Cone Common Market (SCCM or MERCOSUR, acronym in Spanish) in 1991. The same year, the CACM reconstructed its trade agreement with the creation of the Central American Integration System (CAIS), adding Panama and Belize as well as one country from the Caribbean, the Dominican Republic. The Group of Three (G3) was signed by Colombia, Mexico, and Venezuela in 1995 as a free trade area. All of these regional economic integration agreements were understood as part of Latin America's trade policy to diversify their markets and gain from specializations among the region.

Among experiences that have been successful, it is important to mention MERCOSUR. The group has expanded with the inclusion of Venezuela in 2003 and Chile as an observer member. The intra-industry trade between members has grown, particularly for Paraguay, which increased its trade from 30% in 1990 to 50% in 2005. Uruguay's intra-industry trade increased also; more than a third of its trade is made between bloc members. However, Yeats (1997) indicated that trade barriers with third countries have affected the comparative advantages in capital goods due to high trade restriction with nonmembers of the bloc.

In addition, there have been other bilateral agreements with countries outside the region such as Mexico with the United States and Canada (North American Free Trade Agreement [NAFTA]) in 1994, Peru and the United States in 2009, or between regional groups such as MERCOSUR with the European Union (still in the process of signing an agreement).

## Openness Indices

Tariff reductions and regional trade integration agreements were two significant factors for an increase in the participation of external trade (exports plus imports) as a percentage of GDP. This participation is measured as an openness index. It has been used as an indicator of the trade sector's significance in an economy. Figure 83.2 shows the openness indices for selected Latin American countries.

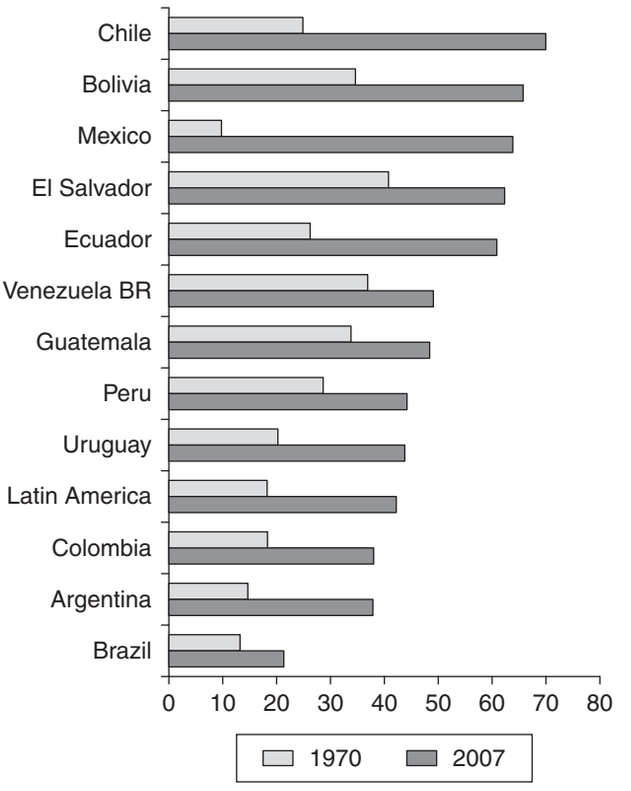


**Figure 83.2** Openness Indices (%)

SOURCE: Based on data from UNCTAD (2008).

Figure 83.2 shows the indices for 2 years, 1970 and 2007. This figure can give an overview of how extensive the external sector is now for Latin American economies. On one hand, the average share of external trade as part of GDP for the region was situated between 10% and 20% in 1970. Some small countries' trade participation in the economy was 40% (El Salvador) or 34% (Guatemala). On the other hand, some countries moved slowly toward opening their economies. In this case, Brazil's trade sector was 13% and Argentina's just 14% of the GDP. Currently, Latin America is one of the most open regions in the world. As can be seen from Figure 83.2, many economies surpassed the 40% mark in 2007. In most cases, it was the increase of exports in the total trade that explained the relative performance.

## Structure of Exports

Trade liberalization had as one of its objectives the increase of manufacturing goods exports as a percentage of total trade. Figure 83.3 shows the structure of exports for Latin America for selected years.

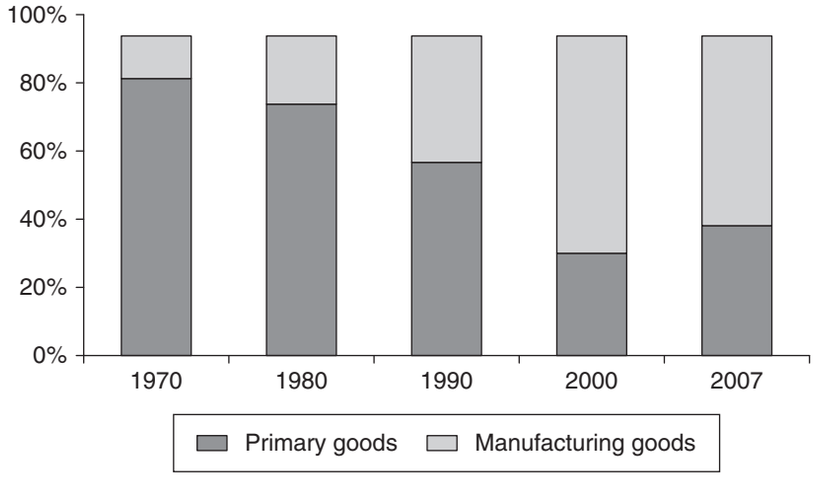


**Figure 83.3** Structure of Exports (%)

SOURCE: Based on data from Economic Commission for Latin America and the Caribbean (ECLAC, 2008).

This figure illustrates how until the 1980s, primary goods were the main export products. The structure of exports started to transform in the 1990s. Manufacturing goods exports were 37% for 1990 and increased to 60% in 2001. Parts and components have played a significant role in increasing manufacturing products, particularly for Brazil and Mexico, as well as exports of finished vehicles. The participation of primary goods began to intensify again in 2003 due to the rise in commodity prices (crude oil was 14.5% of total exports in 2005). In 2006, the exports of manufacturing products were just 48%.

A closer look of leading export products reveals the significance of commodities for the 1970s and the 1980s. In more recent years, there are more manufacturing products exported, particularly those from the automotive industry. The increase of manufacturing goods export has international vertical specialization as one explanation. This trade specialization began in the 1970s with multinational corporations. They established labor-intensive production stages in developing countries to reduce labor costs. The in-bond industry or *maquiladora* is one example of this type of manufacturing production. The maquiladora was a program initiated in 1965 in Mexico along its northern border with the United States. It was backed by a specific tariff scheme between these two countries. It allowed the imports of parts and components from the United States without duties for assembly in Mexico and subsequent export as final product for the U.S. market.

Feenstra (1998) coined the process as integration of trade and disintegration of production. As a result, there is an increase in exports of parts and components, but for countries such as Brazil and Mexico, the export of finished vehicles has become the third most important product.

Recent growth in the demand for commodity products from China and India has caused a rise in prices. On one hand, some Latin American countries (Bolivia, Chile, Peru, and Venezuela) have benefited from this important source of income. The GDP's performance for the 2003–2007 period has as one of its determinants the increase in the demand for commodity products. On the other hand, income revenues from commodity exports have created overvalued pressures on real exchange rates. An overvalued pressure exists when the supply of foreign currency in the economy is greater than the demand. In consequence, the local currency increases in value. An overvalued exchange rate becomes an anti-export bias for noncommodities exports. In addition, inflation pressures have affected the region due in part to the commodity price boom. The average inflation rate for Latin America was 6% for the 2006–2007 period and increased to 9% in 2008. The slowdown in global economic activity due to the international financial crisis will reduce inflationary pressures. The Economic Commission for Latin America and the Caribbean (ECLAC, 2009b) predicted that Latin America's GDP growth rate would fall off to 1.7%.

## Origin and Destination of Exports and Imports

The origin and destination of exports and imports for Latin American countries can be seen in Table 83.2.

**Table 83.2** Exports and Imports Main Partners, 1980–2007

| | *Exports* | | *Imports* | |
|---|---|---|---|---|
| | *1980* | *2007* | *1980* | *2007* |
| United States | 35.4 | 40.3 | 36.9 | 34.7 |
| European Union | 26.8 | 14.3 | 20.6 | 14.8 |
| Latin America | 15.9 | 17.8 | 14.4 | 20.8 |
| China | 0.7 | 5.9 | 0.3 | 7.7 |
| Other countries | 21.3 | 21.7 | 27.8 | 22.0 |

SOURCE: Data from Economic Commission for Latin America and the Caribbean (ECLAC, 2008).

Table 83.2 shows the composition of exports and imports in 1980 and 2007 for trade with selected partners. Regarding exports, the U.S. market has been significant for Latin American exports. In 1980, the region sent 35% of its total exports to the United States, 27 years later increasing that percentage to 40%. The European Union market was important for some manufacturing products (automotive goods) in the 1980s but has gradually lost participation as a destination of Latin American exports (from 27% in 1980 to half of that percentage in 2007). Latin America has also been a destination for exports from members of the region. However, the composition between 1980 and 2007 has not changed significantly (16% in 1980 to 18% in 2007). China has become an important partner for some Latin American countries, particularly for the supply of commodities. Since 2003, China has become the second most important trading partner for Mexico. This performance has meant an increase in its participation as a destination market from 1% in 1980 to 6% in 2007.

The main trade partners for imports of products demanded by Latin American countries can be seen in Table 83.2. The import composition shares similarities with exports. The United States and European Union were significant as sources for imported products. Even though the participation of the European Union has decreased, there are still significant trade ties between both regions. There has been an increase in the import demand from members of the Latin American region from 14% in 1980 to 21% in 2007. Since 2003, China has increased exports to Latin America; so far, the trade balance has been positive for most Latin American countries, but the demand for Chinese products has increased steadily.

## Trade With China

Trade with China has been an important factor for overall trade performance of Latin American economies in the past 5 years. The main goods demanded from this country are primary goods, particularly ores and metals. Figure 83.4 shows the main products imported from Latin America for 1995 and 2006. The economic growth of China since the 1980s has demanded imports of primary products for expanding its industrial base and of some agricultural products that appeal to the more sophisticated Chinese population (e.g., by-products from cattle farming).

Some Latin American countries have consolidated their trade relations with China through bilateral trade agreements. Chile was the first Latin American country to sign a free trade agreement (FTA) with China in 2005. Peru initiated negotiation toward an FTA the same year, signing an agreement in 2008. Rosales and Kuwuyama (2007) pointed out that the benefits of trading with China have been uneven for the region. South American countries have gained from an increase in terms of trade (export prices/import prices), while Central American countries have been hurt by the competition of Chinese products in the U.S. manufacturing market.

Mexico has seen its share of manufactured products decrease in the U.S. market (textiles and apparel) due to an increase in imports from China. Chiquiar, Fragoso, and Ramos-Francia (2007) calculated that Mexico lost comparative advantages (Mexican exports vs. world exports to the U.S. market) in the textile and apparel sectors because of competition by Chinese products. Even though Mexico enjoyed free-market access to the U.S. market with NAFTA, the low cost of Chinese manufacturing as well as changes in the WTO regulations explain the increase of

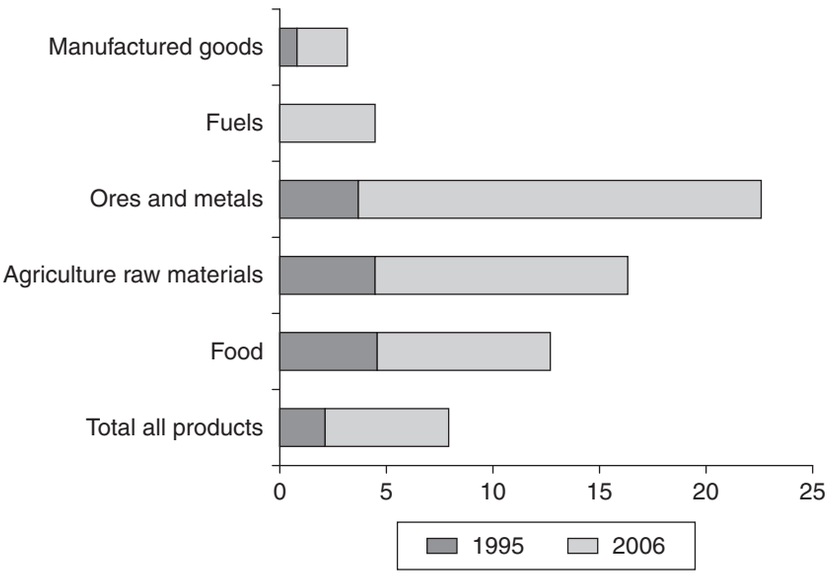


**Figure 83.4** China: Main Products Imported From Latin America (%)

SOURCE: Based on data from United Nations Conference on Trade and Development (UNCTAD, 2008).

Chinese goods in the U.S. market. Gallagher, Moreno-Brid, and Porzecanski (2008) stated that Mexico has lost competitiveness in 15 non-oil-exported products in the U.S. market due to Chinese competition. Their measurement indicates that industries that depend on unskilled labor are the most threatened.

## Effects on GDP

Finally, we analyze the possible links between trade performance and its effects on GDP, employment. and industrial production. There is some argument about the positive effects of trade performance on economic growth. Empirical evidence for the 1990s has not been conclusive about this relationship (Giles & Williams, 2000). Furthermore, one of the most debated issues at that time was the slow economic growth that accompanied the posttrade liberalization period. Latin America's per capita GDP growth rate for 1990–2004 was 0.9%, a much lower rate than the period of ISI (2.6% for 1950–1980). This rate was lower than for the East Asian countries for the 1990–2000 period (3.95%) and for the world (1.2%).

Ffrench-Davis (2005) reviewed the different approaches to explain the low economic growth of the 1990s. Some of his focus was on the stabilization programs during the 1980s and the lack of complementary reforms to support manufactured exports. For example, Dijkstra (2000) affirmed that trade liberalization effects on industrial development were mixed depending on the level of development of the industrial base in Latin American countries. He found that industrial structural changes in Chile and Brazil were explained by domestic demand and exchange rate factors rather than an increase in exports. Other studies criticized the Washington Consensus reforms and their failure to deliver the expected benefits of an open economy. This disappointment with the neoliberal economic reforms has been one of the reasons for recent trends toward more state intervention in the economy. Countries such as Venezuela, Bolivia, and Ecuador have distanced themselves from the Washington Consensus policies.

Figure 83.5 shows some indicators of growth, per capita GDP, and exports for the 2000–2008 period for Latin America.

The GDP per capita and the value-added manufacturing and total export rates of growth are depicted in the left axis, and the urban open unemployment rate is depicted in the right axis. As can be seen in the figure, the per capita GDP has grown steadily since 2003, with percentages that the region has not experienced since the ISI period (around 5%). The performance of total exports varied over this period, since between 2001 and 2003, there was a decline in the rate of growth, but it has started to increase since 2003 due to an important increase in primary product exports (more than 15% since 2004). Besides, it is important to take into consideration that during this same period, primary product exports explained more than 10% of the total GDP. As mentioned before, the international demand for these products

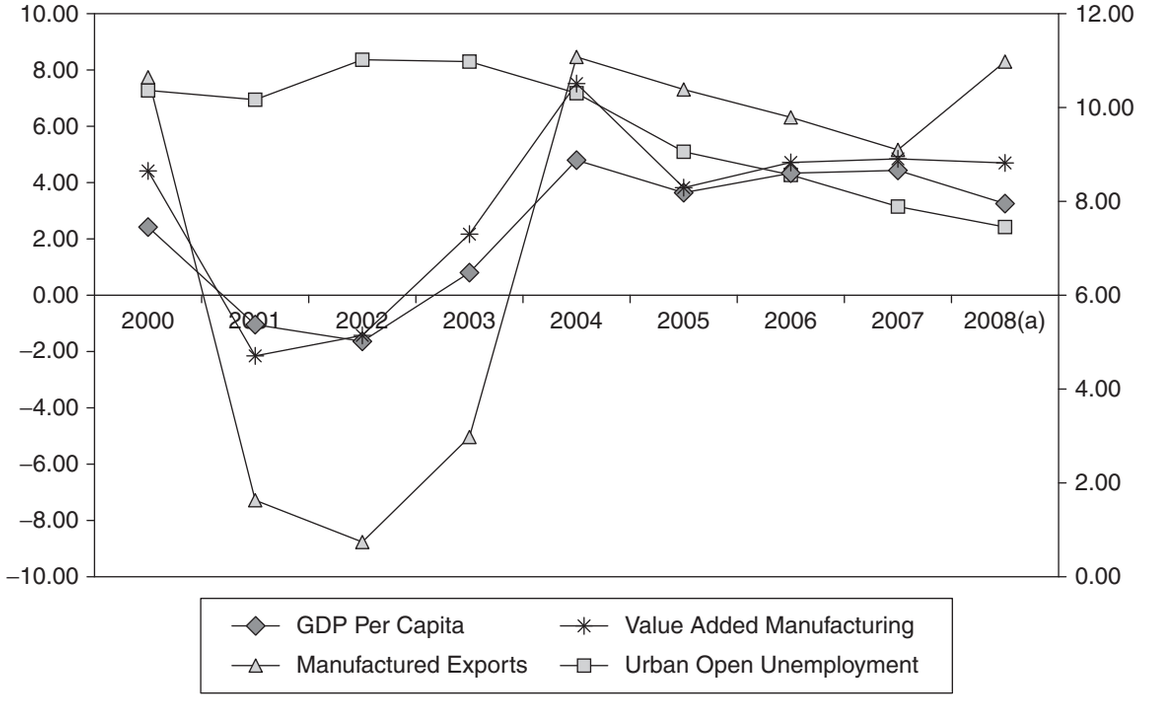


**Figure 83.5** Exports Growth Rates and Economic Performance (%)

SOURCES: Based on data from United Nations Conference on Trade and Development (UNCTAD, 2008) and Economic Commission for Latin America and the Caribbean (ECLAC, 2008).
NOTE: (a) Estimated figures.

and the favorable terms of trade have helped to shape these results. Since 2004, the value added for manufacturing products grew at rates close to 5% until 2008. One of the results of the positive performance of the economies can be seen in a reduction of unemployment. From 8% in 2000, it was reduced to close to 5% in 2008. However, the international financial crisis of 2008 affected the economic performance of the following years for the region, as we will discuss in the next section.

## Financial Crisis of 2008 and Trade Performance

Latin American countries began to experience the slowdown of the global economy in 2008, particularly during the last quarter of the same year, when major problems in the U.S. financial crisis could not be solved and spread over European and Japanese financial markets. This financial crisis was a result of the real estate market's collapse and the damage to the related financial system. The immediate consequence was a reduction in consumer spending, which brought an increase in unemployment. According to the National Bureau of Economic Research (NBER), the U.S. economy officially entered into a period of recession on December 1, 2007 (NBER, 2008).

Dealing with financial crises is not new for Latin American economies. For example, the Great Depression of the 1930s and the debt crisis of the 1980s had lasting negative consequences for many economies and took several years to stabilize. For Latin America, the main transmission channels of the global economic crisis could come from the reduction in export demand, decline in prices of commodity goods, and therefore a reduction in terms of trade as well as remittances and restrictions in financial credit sources (World Bank, 2008).

According to a number of analyses (ECLAC, 2008; International Development Bank [IDB], 2008; World Bank, 2008), most of Latin America is better prepared than before to confront external shocks due to improved management of macroeconomic policies. During the 1990s and, to a major extent, in the beginning of the twenty-first century, these countries have reduced their external debt, increased FDI flows, limited fiscal spending, and followed semi-flexible exchange rates. These macroeconomic measures should reduce the possibility of an economic crisis of big proportions. Also, the decline in foreign demand would reduce inflationary pressures that increase external demand caused during 2008 in some countries of the region such as Peru, Argentina, or Bolivia. In general terms, ECLAC (2009a) has projected a negative growth of –0.3% for Latin America and the Caribbean for 2009 and a decline of 15% in the region's terms of trade.

The effects of the financial crisis will not be homogeneous in the region. First, it is expected that countries with commercial ties to the U.S. market will experience a sharp decline in demand for their exports. That is the case of Mexico, in which more than 60% of total exports go to the U.S. market. The Mexican economy also depends on the flow of remittances from Mexican workers living in the United States, so if those decrease, that could worsen the trade account. A reduction in remittances was expected for the rest of 2009. It could have an impact on the Mexican GDP since the remittances account for 3% of the total GDP. Moreover, ECLAC estimates a reduction of –2.0% in total GDP in 2009 (ECLAC, 2009a). The Central American countries face similar economic perspectives. The U.S. market is significant for Central America's maquiladora products; a reduction in demand would affect these economies. Remittances also are significant for countries such as El Salvador (30% of GDP) and Honduras (20% of GDP).

The economies of South America are more diversified in terms of export destinations. Even though the U.S. market is considerable for their products, there is an important growing trade with Asian countries and intra-regional trade with countries in South America. Since the increase in demand for commodity products originates in Asian countries, particularly from China, the fall of commodity prices will have an effect on volume of exports. However, the export performance predictions for South America are not worrisome (with the exception of Argentina) since it is expected that demand of goods from countries such as China and India will continue, and these countries will maintain economic ties with South America to secure sources of primary goods for their own economic growth, although not at the same pace as during the past 5 years.

## Conclusion

Trade liberalization began during the 1990s for most Latin American countries. After decades of inward-looking economic policies and the struggles to develop a strong industrial base, countries such as Brazil, Mexico, and Argentina are considered today to be emerging economies with competitive industrial sectors.

However, the economic reforms of the 1990s have proven to be a challenge for the region since the eruption of several macroeconomic collapses such as in Argentina in 2001. For that reason, there have been criticisms of the Washington Consensus approach. In fact, the discontent with the market fundamentalism of the Washington Consensus has been the reason for the conception of a new path of development in Latin America. Under a different approach, the state again is taking a role in the economy, as in Venezuela, Bolivia, and Ecuador and, to a lesser measure, in Brazil and Argentina.

In general, the first years of the twenty-first century have been relatively successful for Latin American countries. After almost two decades of stabilization programs, the region has started to experience steady economic growth from 2003 until 2007. The impact on GDP growth has been considerable, around 5% per year for the same period, a percentage that the region has not seen since three

decades earlier. This performance has been explained by the rapid increase in primary goods prices. However, economic reforms from the 1990s have been a factor in the development of its external trade sector.

One of the results of trade liberalization has been a transformation in the composition of exports, with an increase of manufactured goods as a share of total exports. However, trade performance has not been homogeneous for the whole region. While Mexico's exports are concentrated in manufactured products (including products from the maquiladora program) directed to the U.S. market, countries from South America are more diversified in terms of their export destination and share a significant intraregional trade. MERCOSUR plays an important role for the region as well as the Andean Community.

The rise of China in the global economy has been a turning point for Latin American primary goods. Since 2003, these countries have become one of the main origins of exports as well as destinations for imports coming from China. In industrial sectors such as textiles, China has been a serious competitor with Mexico and Central American countries in the U.S. market.

The financial global crisis of 2008 will have an important impact on Latin American economies. At the beginning of 2008, the economic growth rate prediction for the region was around 3%. At the beginning of 2009, that percentage was dropped to –0.1%. Latin America faces the current global financial crisis as a new test of how it can manage an external shock without entering a recession (such as in the 1930s or the 1980s). The crisis will hit countries that depend on the U.S. market for their exports more severely, such as in Mexico and Central America. Moreover, incomes from remittances also will diminish during this period. Meanwhile, South America will experience the impact of financial restrictions and the decline in the demand for primary goods. The crisis will also help countries such as Peru to cope with imminent inflation problems.

In the second era of globalization, Latin American countries are much better prepared to face external shocks. The region has stronger institutions with significant international reserves as well as fiscal and monetary discipline that will help to ease negative impacts. The financial crisis has been analyzed as evidence of market failures. For this reason, Latin America is taking this time to reassess, searching for a new path of economic development. It includes the structural economic characteristics without the need for protectionist policies or inward-looking economic policies.

The global financial crisis can be seen as an opportunity for a change in trade pattern specialization. First, a diversification of the destination of exports can reduce the dependence of demand of one country (i.e., the U.S. market). Second, increasing intra-industry trade between members of the region and efforts toward regional integration can increase the trade in manufacturing goods. Third, even though Latin American countries have a comparative advantage in primary products, there are industrial sectors that have been proven successful in the international arena, such as the automotive industry in Brazil and Mexico. Therefore, Latin America could increase the share of manufacturing products in total trade and reduce the volatilities of prices of primary goods. All of the above could result in a new role in the international economy for the region.

## References and Further Readings


Balassa, B. (1971). Regional integration and trade liberalization in Latin America. *Journal of Common Market Studies, 10,* 58–78.

Bulmer-Thomas, V. (2003). *The economic history of Latin America since independence* (2nd ed.). Cambridge, UK: Cambridge University Press.

Cardoso, E., & Helwege, A. (1997). *Latin America's economy: Diversity, trends and conflicts.* Cambridge: MIT Press.

Chiquiar, D., Fragoso, E., & Ramos-Francia, M. (2007). *La ventaja comparativa y el desempeño de las exportaciones manufactureras mexicanas en el periodo 1996–2005* [Comparative advantage and the performance of the Mexican manufactured export during 1995–2005 period] (Working Paper 2007–12). Mexico DF: Banco de Mexico.

Dijkstra, G. (2000). Trade liberalization and industrial development in Latin America. *World Development, 28,* 1567–1582.

Dornbusch, R. (1992). The case for trade liberalization in developing countries. *Journal of Economic Perspectives, 6,* 69–85.

Economic Commission for Latin America and the Caribbean (ECLAC). (2008). *Latin America and the Caribbean in the world economy, 2007: Trends 2008.* Santiago de Chile: Author.

Economic Commission for Latin America and the Caribbean (ECLAC). (2009a, April). *Economic growth in Latin America and the Caribbean will fall to –0.3% in 2009* (Press release). Santiago de Chile: Author.

Economic Commission for Latin America and the Caribbean (ECLAC). (2009b, December). *El comercio internacional en América Latina y el Caribe en 2009. Crisis y recuperación*. Santiago de Chile: Author.

Edwards, S. (2004). *Crisis and reform in Latin America: From despair to hope.* Washington, DC: World Bank.

Feenstra, R. C. (1998). Integration of trade and disintegration of production. *Journal of Economic Perspectives, 12*(4), 31–50.

Ffrench-Davis, R. (2005). *Reforming Latin America's economies: After market fundamentalism.* New York City: Palgrave-Macmillan.

Flam, H., & Grunwald, J. (1985). *The global factory: Foreign assembly in international trade.* Washington, DC: The Brookings Institution.

Franco, P. (2007). *The puzzle of Latin American economic development.* Lanham, MD: Rowman & Littlefield.

Frieden, J., Pastor, M., Jr., & Tomz, M. (Eds.). (2000). *Modern political economy and Latin America: Theory and policy.* Boulder, CO: Westview.

Gallagher, K., Moreno-Brid, J. C., & Porzecanski, R. (2008). The dynamism of Mexican exports: Lost in (Chinese) translation? *World Development, 36,* 1365–1380.

Giles, J., & Williams, C. (2000). Export-led growth: A survey of the empirical literature and some non-causality results. *Journal of International Trade & Economic Development, 9,* 261–337.

Gwartney, J., & Lawson, R. (with Norton, S.). (2008). *Economic freedom of the world: 2008 annual report.* Vancouver, British Columbia, Canada: The Fraser Institute.

Helwege, A. (1995). Poverty in Latin America: Back to the abyss? *Journal of Interamerican Studies & World Affairs, 37*(3), 99–123.

Hofman, A. (2000). *The economic development of Latin America in the twentieth century.* Northampton, MA: Edward Elgar.

International Development Bank. (2008, October). *The fiscal impact of the international financial crisis on Latin America and the Caribbean* (Technical note). Washington, DC: Author.

Krugman, P., & Obstfeld, M. (2000). *International economics: Theory and policy* (5th ed.). Reading, MA: Addison-Wesley Longman.

Lederman, D., Olarreaga, M., & Perry, G. (Eds.). (2009). *China's and India's challenge to Latin America: Opportunity of threat.* Washington, DC: World Bank.

Lora, E. (2001). *Structural reforms in Latin America: What has been reformed and how to measure it* (Research Department Working Paper Series No. 466). Washington, DC: Inter-American Development Bank.

Lustig, N. (1994). *México. Hacia la Reconstrución de una Economía* [Mexico: The remaking of an economy]. México DF: El Colegio de México, Fondo de Cultura Económica.

National Bureau of Economic Research (NBER). (2008). *Determination of the December 2007 peak in economic activity.* Cambridge, MA: Business Cycle Dating Committee, NBER.

Rodrik, D. (1999). *The new global economy and developing countries: Making openness work* (Policy Essay No. 24). Washington, DC: Overseas Development Council.

Rosales, O., & Kuwuyama, M. (2007). Latin America meets China and India: Prospects and challenges for trade and investment. *Cepal Review, 93,* 81–103.

Santiso, J. E. (2007). *The visible hand of China in Latin America.* Paris: OECD.

United Nations Conference on Trade and Development (UNCTAD). (2008). *Handbook of statistics 2008.* New York: Author. Retrieved from http://stats.unctad.org/Handbook/TableViewer/tableView.aspx

United Nations Development Programme (UNDP). (1990). *Human development report 1990: Concept and measurement of human development, UNDP.* Oxford, UK: Oxford University Press.

Williamson, J. (1990). What Washington means by policy reform. In J. Willianson (Ed.), *Latin American adjustment: How much has happened?* (pp. 7–20). Washington, DC: Institute for International Economics.

World Bank. (2008). *Global economic prospects: Commodities at crossroads 2009.* Washington, DC: Author.

Yeats, A. (1997). *Does Mercosur's trade performance raise concerns about the regional trade arrangements?* (World Bank Policy Research Working Paper 1729). Washington, DC: World Bank.

# PART VII

## Emerging Areas in Economics

# 84

# Behavioral Economics

Nathan Berg
*University of Texas–Dallas*

Behavioral economics is the subfield of economics that borrows from psychology, empirically tests assumptions used elsewhere in economics, and provides theories that aim to be more realistic and closely tied to experimental and field data. In a frequently cited survey article, Rabin (1998) describes behavioral economics as "psychology and economics," which is a frequently used synonym for behavioral economics. Similarly, Camerer (1999) defines behavioral economics as a research program aimed at reunifying psychology and economics.

## Definitions and Naming Problems

Reunification is a relevant description because of the rather tumultuous relationship between psychology and economics in the arc of economic history. A number of preeminent founders of important schools of economic thought, including Adam Smith, wrote extensively on psychological dimensions of human experience and economic behavior, while later economists sometimes sought explicitly to exclude psychology from economic analysis. For example, Slutsky (1915/1952), whose famous equation is taught to nearly all upper-level microeconomics students, sought to erect a boundary excluding psychology from economics: "If we wish to place economic science upon a solid basis, we must make it completely independent of psychological assumptions" (p. 27).

Although historical accounts vary, one standard narrative holds that in the twentieth century, neoclassical economists made an intentional break with psychology in contrast to earlier classical and institutional economists who actively integrated psychology into their writings on economics (e.g., Bruni & Sugden, 2007). In twentieth-century economics' break with psychology, one especially important source is Milton Friedman's (1953) essay. In it, Friedman argues that unrealistic or even obviously untrue assumptions—especially, the core assumption used throughout much of contemporary economics (including much of behavioral economics) that all behavior can be modeled as resulting from decision makers solving constrained optimization problems—are perfectly legitimate, so long as they produce accurate predictions. Friedman put forth the analogy of a billiards player selecting shots "as if" he or she were solving a set of equations describing the paths of billiards balls based on Newtonian physics. We know that most expert billiards players have not studied academic physics and therefore do not in fact solve a set of equations each time they set up a shot. Nevertheless, Friedman argues that this model based on manifestly wrong assumptions should be judged strictly in terms of the predictions it makes and not the realism of its assumptions.

In contrast to Friedman's professed lack of interest in investigating the realism of assumptions, behavioral economists have made it a core theme in their work to empirically test assumptions in economic models and modify theory according to the results they observe. Despite this difference with neoclassical economists such as Friedman, behavioral economists frequently use as-if arguments to defend behavioral models, leading some methodological observers to see more similarity than contrast in the behavioral and neoclassical approaches (Berg & Gigerenzer, in press).

## Bounded Rationality

The term *bounded rationality,* coined by Nobel laureate Herbert Simon (1986), is strongly associated with behavioral economics, although there appears to be far less agreement on the term's meaning. The neoclassical model assumes that economic man, or *homo economicus,* is infinitely self-interested, infinitely capable of processing information and solving optimization problems, and infinitely self-disciplined or self-consistent when it comes to having the willpower to execute one's plans—whether those plans concern how much junk food to eat or how much to save for retirement. In contrast, much of behavioral economics focuses on limits, or bounds, on one or more of these three assumptions. Thus, bounded self-interest, bounded information-processing capacity, and bounded willpower are three guiding themes in the behavioral economics literature.

Bounded self-interest enjoys widespread appeal in behavioral economics, which has proposed numerous models of so-called social preferences to address a number of observations from human experiments that appear to falsify the assumption that people maximize their own monetary payoffs. A decision maker with social preferences cares about the material or monetary payoffs of others as well as his or her own, although the manner in which concern for others' payoffs is expressed can take a variety of forms. For example, a person with social preferences might be happier when others are worse off, which is sometimes described as spite; be happier when others are better off, which is sometimes described as altruism; prefer equal over unequal allocations of money, which is sometimes described as inequality aversion; prefer allocations in which the sum of all people's payoffs is maximized, which is sometimes described as a preference for social welfare; prefer allocations in which the least well-off person has a larger payoff, which is sometimes described as a Rawlsian preference; or prefer allocations of resources in which his or her payoff is large relative to others, which is sometimes described as a competitive preference (Charness & Grosskopf, 2001). Common to all these variations and many other forms of social preferences is that people are not generally indifferent between two allocations of payoffs for all members in a group just because their own monetary payoff is the same. This violates the common neoclassical assumption that people are infinitely self-interested because it would imply that people are indifferent so long as their own material payoffs are held constant.

Bounded information-processing capacity is another active area within behavioral economics, which would have looked very much out of place in the mainstream economics literature only three decades ago. Topics in this area include limited memory, limited attention, limited number of degrees of perspective taking in strategic interaction, limited perceptual capacity, distorted beliefs, and decision and inference processes that violate various tenets of logic and probability theory (see Camerer, 2003, for examples).

Bounded willpower, often described as time inconsistency or dynamic inconsistency, is another large and growing part of the behavioral economics research program. In the neoclassical optimization model, decision makers choose a sequence of actions through time by selecting the best feasible sequence, with virtually no mention of the costs associated with implementing that plan over the course of people's lives. If there is no new information, then the neoclassical intertemporal choice problem is decided once and for all before the first action in the sequence is taken. Unlike the neoclassical model's assumption that acting on the optimal plan of action through time is costless, behavioral economists studying bounded willpower focus squarely on the tension between what a person wants himself or herself to do tomorrow versus what he or she actually does. This tension can be described as subjective inconsistency concerning what is best for oneself at a specific point in time, which, contrary to the neoclassical assumption, changes as a function of the time at which the decision is considered. One can think of planning now to start working on a term paper tomorrow but then tomorrow deciding to do something else instead—and regretting it after the fact.

## Empirical Realism

Proponents of bringing psychology more deeply into economics argue that it is necessary to depart from the assumptions of *homo economicus* to achieve improved empirical realism (i.e., more accurate descriptions of economic behavior and better predictions following a change in policy or other economic conditions). In an article published in the *Journal of Business* titled "Rationality in Psychology and Economics," Simon (1986) writes,

> The substantive theories of rationality that are held by neoclassical economists lack an empirically based theory of choice. Procedural theories of rationality, which attempt to explain what information people use when making choices and how information is processed, could greatly improve the descriptive and forecasting ability of economic analysis. (p. S209)

Improving the empirical realism of economic analysis is a primary and ongoing motivation in behavioral economics, frequently stated in the writings and presentations of behavioral economists.

## Example of Reference Point–Dependent Utility Functions

Similarly to Simon, Rabin (1998) argues that theories and experimental results from psychology enrich mainstream economics. But Rabin's idea about how behavioral economists can bring more empirical realism into economics is much more narrowly circumscribed than Simon's, with Rabin essentially arguing that behavioral

economics should proceed within the utility maximization model of neoclassical economics. Despite their occasional claims to radical or revolutionary methodological innovation, many behavioral economists side with neoclassical economists in viewing constrained optimization as a nonnegotiable methodological tenet that defines and distinguishes economics from other disciplines. Rabin says that the new empirical content that behavioral economists bring to bear will help economics as a whole to more realistically describe people's utility functions.

For example, as mentioned earlier in the discussion of the neoclassical model's assumption of unbounded self-interest, this assumption is often interpreted to mean that consumers care only about their own levels of consumption and that workers care only about their own income, irrespective of what others are consuming or earning. Behavioral economics models, in contrast, allow utility to depend on the *difference* between one's own level of consumption or income and a reference point level. The reference point level might reflect what one is accustomed to or reflect a social comparison made with respect to the average level within a social group.

Thus, a worker with a behavioral reference point–dependent utility function might prefer an annual salary of $90,000 at a company where the average worker earns $50,000 over a salary of $95,000 at a company where the average worker earns $200,000. In the standard economic model, only the worker's own payoffs should determine the ranking of job opportunities, holding all else equal, and not the comparison of one's own income with that of other workers. The reference point–dependent utility function tries to reflect the observation that many normal, healthy, and socially intelligent people do in fact care about their own payoffs relative to others. For some workers, it may be worthwhile to trade off a few thousand dollars of their own salary for a work environment where the relative pay structure is more to their liking (e.g., a feeling of relative high status in the $90,000 job being subjectively worth more than the extra $5,000 of income at the $95,000 job). It should be mentioned, however, that reference point–dependent theories in behavioral economics are not entirely new. One finds interpersonal comparisons in Veblen's (1899/1994) concept of conspicuous consumption from his classic *The Theory of the Working Class* and even earlier among some classical economists.

## Debates About the Realism of Assumptions in Economic Models

There is active debate between behavioral and nonbehavioral economists—and among behavioral economists—about the extent to which empirical realism is being achieved by the behavioral economics research program. These two distinct layers of debate need to be untangled to appreciate the different issues at play and how behavioral economics is likely to influence public policy now that the Obama administration has recruited among its top advisers a number of behavioral economists, including Richard Thaler, Cass Sunstein, and Daniel Kahneman.

When trying to convince neoclassical economists who are skeptical about the need for behavioral economics, behavioral economists point to the improved ability of their psychology-inspired models to fit data collected from a variety of sources, including experimental, macroeconomic, and financial market data. Skeptics from outside behavioral economics have questioned whether the deviations from neoclassical assumptions have any important consequences for the economy as a whole, suggesting that they might perhaps "average out" in the aggregate. Skepticism about the relevance of experimental data remains strong, with many doubts expressed about whether the college students who participate in economic experiments can be relied upon to teach us anything new about economics and whether anything learned in one laboratory experiment can be generalized to broader populations in the economy—the so-called problem of external validity. Experimentalists have responded that the reason they carefully incentivize decisions by making subject payments dependent on their decisions is to make it costly for them to misrepresent their true preferences. Experimentalists have addressed the issue of external validity by going into the field with so-called field experiments and by conducting experiments among different subpopulations, such as financial market traders, Japanese fishermen, and other groups of adult workers (e.g., Carpenter & Seki, 2006).

Within behavioral economics, a different debate takes place. Among behavioral economists, despite a shared commitment to borrowing from psychology and other disciplines, there remains tension over how far to move away from constrained optimization as the singular organizing framework of neoclassical theory and in much of behavioral economics, too. An alternative approach, advocated by a minority of more psychology- and less economics-inspired behavioral economists, seeks to break more substantially with neoclassical economics, dispensing with optimization theory as a necessary step in deriving equations that describe behavior. Constrained optimization, whether in behavioral or neoclassical economics, assumes that decision makers see a well-defined choice set; exhaustively scan this set, plugging each possible action into a scalar-valued objective function, which might include parameters intended to capture psychological phenomena; weigh the costs and benefits associated with each action, which includes psychic costs and benefits; and finally choose the element in the choice set with the highest value according to the objective function. There is very little direct evidence of people making decisions—especially high-stakes decisions, such as choosing a career, buying a house, or choosing whom to marry—according to the constrained optimization process just described. In many real-world decisions such as those just mentioned, the choice set is impossibly large to clearly define and exhaustively search through. In other settings such as choosing a life partner or whom to marry, constrained optimization would be seen by some to violate important social norms.

Instead, critics such as Gigerenzer and Selten (2001) attempt to base theory directly on empirical description of actual decision processes. Like other economists, these critics use equations to describe behavior. However, their behavioral equations skip the step of deriving behavioral equations as solutions to constrained optimization problems. To these researchers, theorizing and observing how decision makers deal with the overwhelmingly high-dimensional choice sets they face, quickly searching for a good-enough action and discarding the rest, is a fundamental scientific question of primary importance. Herbert Simon (1986) referred to such threshold-seeking behavior as *satisficing* as distinct from *optimizing.*

### Methodological Pluralism

Another theme in behavioral economics derives from its willingness to borrow from psychology and other disciplines such as sociology, biology, and neuroscience. To appreciate why methodological pluralism is characteristic of behavioral economics, one should recall that in neoclassical economics, there is a singular behavioral model applied to all problems as well as a number of prominent efforts in economic history to expunge influence from other social sciences such as psychology and sociology. Although the structure of choice sets and the objective functions change depending on the application, contemporary economists typically apply the maximization principle to virtually every decision problem they consider. Consumer choice is modeled as utility maximization, firm behavior is modeled as profit maximization, and the evaluation of public policy is analyzed via a social welfare function whose maximized value depends systematically on parameters representing policy tools. In contrast, commitment to improved empirical description and its normative application to policy problems motivates behavioral economists, in many cases, to draw on a wider set of methodological tools, although the breadth of this pluralism is a matter of debate, as indicated in the previous section.

### Naming Problems

The term *behavioral economics* is often associated with the pioneering work of George Katona (1951). Behavioral economists sometimes joke that the name of their subfield is redundant since economics is a social science in which the objects of study depend directly on human behavior. "Isn't all economics supposed to be about behavior?" the quip goes. Despite the appearance of a "distinction with no distinction" inherent in its name, proponents of behavioral economics argue that there is good reason for the explicit emphasis on accurate description of human behavior, as indicated by the word *behavioral* in behavioral economics.

In psychology, there is a sharp distinction between the terms *behaviorist* and *behavioral.* Behaviorism refers to research and researchers that draw on the work of B. F. Skinner in hypothesizing that most behavior can be explained in terms of adaptation to past rewards and punishments. Behaviorism rejects investigation of mental states or other psychic determinants of behavior. Thus, behaviorism is more similar to neoclassical economics because both schools of thought rely on a singular story about what underlies observed behavior while expressing overt antipathy toward the inclusion of mental states, cognitive processing, or emotion in their models. One frequently finds mistaken references to behavioral economists as "behaviorists" in the popular press, whereas "behavioralists" would be more accurate.

### Behavioral Economics and Experimental Economics

Strong connections between behavioral and experimental economics can be seen in behavioral economists' reliance on experimental data to test assumptions and motivate new theoretical models. There nevertheless remains a distinction to be made (Camerer & Loewenstein, 2004). Some experimental economists do not identify with behavioral economics at all but rather place their work firmly within the rational choice category, studying, for example, the performance of different market institutions and factors that enhance the predictions of neoclassical theory. Experimental economics is defined by the method of experimentation, whereas behavioral economics is methodologically eclectic. The two subfields have subtly different standards about proper technique for conducting lab experiments and very different interests about the kinds of data that are most interesting to collect. Therefore, it is incorrect to automatically place experimental work under the heading of behavioral economics. In the other direction, there are many behavioral economists working on theoretical problems or using nonexperimental data. Thus, although behavioral and experimental economists frequently work complementarily on related sets of issues, there are strong networks of researchers working in the disjoint subsets of these subfields as well.

## Frequently Discussed Violations of Internally Consistent Logic

This section describes several well-known violations of the rational choice model based on reasoning that allegedly suffers from internal inconsistency. The following example about deciding where to buy a textbook illustrates the kind of inconsistencies that are frequently studied in behavioral economics. Readers are encouraged to decide for themselves how reasonable or unreasonable these inconsistencies in fact are.

Suppose you are shopping for a required textbook. A bookstore across the street from where you work sells the book for $80. Another bookstore, which is 15 minutes

away by car or public transportation, sells the book for only $45. Which do you choose: (A) Buy the book at the nearby store for $80, or (B) Buy the book at the farther away store for $45?

Just as standard theory does not prescribe whether it is better to spend your money buying apples versus oranges, so, too, standard economic theory takes no stand on which choice of stores is correct or rational. But now consider a second choice problem.

Suppose you are buying a plane ticket to Europe. The travel agent across the street from where you work sells the ticket for $1,120. Another travel agency, which is 15 minutes away by car or public transportation, sells the same ticket for $1,085. Which do you choose: (C) Buy the ticket from the nearby agency for $1,120, or (D) Buy the ticket at the farther away store for $1,085?

Considered in isolation, either A or B is consistent with rationality in the first choice problem, and either C or D can be rationalized in the second choice problem—as long as these problems are considered alone. Internal consistency requires, however, that a rational person choosing A in the first problem must choose C in the second problem and that a rational person choosing B in the first problem must choose D in the second problem.

Based on extensive data from pairs of choices like the ones just described, many people prefer B in the first problem (i.e., the $35 saved on the cheaper textbook justifies spending extra time and money on the 30-minute round-trip commute) while preferring C in the second problem (i.e., the $35 saved on the cheaper airline ticket does not justify spending extra time and money on the 30-minute round-trip commute). According to axiomatic rationality, this pair of choices is inconsistent and therefore irrational. Yet many competent and successful people, without any obvious symptoms of economic pathology, choose this very combination of allegedly irrational (i.e., inconsistent) decisions.

One explanation is that some people weigh the $35 savings in percentage terms relative to total price. An $80 textbook is more than 75% more expensive than a $45 textbook, whereas a $1,120 planet ticket is less than 5% more expensive than a $1,085 ticket. Nevertheless, the logic of the cost-benefit model of human behavior at the core of rational choice, or neoclassical, economics regards dollars saved—and not percentages saved—as the relevant data.

Choosing A over B, in the eyes of a neoclassical economist, reveals an algebraic inequality:

utility of saving $35 > disutility of a 30-minute round-trip commute.

Choosing D over C reveals another algebraic inequality:

utility of saving $35 < disutility of a 30-minute round-trip commute.

Thus, choosing B over A and C over D leads to inconsistent inequalities, which violate the axiomatic definition of a rational preference ordering. One may justifiably ask, so what? Skepticism over the importance of such violations of axiomatic rationality is discussed in a subsequent section under the heading "Rationality."

## Endowment Effect

Suppose you walk into a music store looking for a guitar. The very cheap ones do not produce a sound you like. And most of the guitars with beautiful sounds are priced thousands outside your budget. You finally find one that has a nice sound and a moderate price of $800. Given the guitar's qualities and its price, you are almost indifferent between owning the guitar and parting with $800, on one hand, versus not owning it and hanging on to your money, on the other. You go ahead and buy the guitar. After bringing it home, enjoying playing it, and generally feeling satisfied with your purchase, you receive a phone call the very next day from the music store asking if you would sell the guitar back. The store offers $1,000, giving you an extra $200 for your trouble. Would you sell it back?

According to the standard cost-benefit theory, if you were indifferent between the guitar and $800, then you should be more than happy to sell it back for anything over $800—as long as the amount extra includes enough to compensate for the hassle, time, and transport costs of returning it to the store (and also assuming you haven't run into someone else who wants to buy the guitar and is willing to pay a higher price). Hoping to bargain for a higher offer from the music store, you might demand something far above $800 at first. But after bargaining, when facing a credible take-it-or-leave-it last offer, anything that gives you $800 plus compensation for returning to the store should leave you better off than holding onto the guitar.

Based on data showing the prevalence of the endowment effect, however, behavioral economists would predict that you probably will choose to hang onto the guitar even if the guitar store's offer climbed well over $1,000. The endowment effect occurs whenever owning something shifts the price at which one is willing to sell it upward to a significantly higher level than the price at which the same person is willing to buy it. In the neoclassical theory taught in undergraduate textbooks with demand curves and indifference curves, an important maintained assumption that is not frequently discussed in much depth is that, for small changes in a consumer's consumption bundle, the amount of money needed to just compensate for a reduction in consumption is exactly equal to the consumer's willingness to pay to acquire that same change in consumption. In a related experiment, Carmon and Ariely (2000) showed that Duke University students who win the right to buy sports tickets in a university lottery valued these tickets—which they became the owner of by chance—roughly 14 times as much as students who had

entered the lottery but did not win. Some researchers have linked the endowment effect to loss aversion, which refers to the phenomenon by which the psychic pain of parting with an object one currently owns is greater than the psychic gain from acquiring it. Thus, rather than ownership shifting the pleasure derived from a good or service upward, some experimental evidence suggests that an increase in pain at dispossessing oneself of a good or service generates the gap by which willingness to accept is significantly higher than willingness to pay.

### Preference Reversals

Lichtenstein and Slovic (1971) and Thaler and Tversky (1990) produced evidence that shook many observers' confidence in a fundamental economic concept—the preference ordering. These and other authors' observed preference reversals, which call into question the very existence of stable preferences that neoclassical analysis depends on, occurred in a variety of contexts: gamblers' valuations of risky gambles at casinos, citizens' valuations of public policies aimed at saving lives, firms' evaluations of job applicants, consumers' feelings toward everyday consumer products, and savers' attitudes toward different savings plans.

In a typical experiment, a group of subjects is asked to choose one of two gambles: (A) win $4 with probability 8/9, or (B) win $40 with probability 1/9. When asked to choose between the two, the less risky gamble A, which provides a high probability of winning a small amount, is typically chosen over B, which pays slightly more on average but pays off zero most of the time. Next, another group of experimental subjects is asked to assign a dollar value to both gambles, A and B, stating the amount of money they would be willing to pay for A and B, respectively. Most subjects typically choose A over B when asked to *choose.* But most subjects place a larger dollar valuation on gamble B when asked to *evaluate* in terms of money. Choosing A over B, while valuing B more highly than A in dollar terms, is a preference reversal. In neoclassical theory, a person with a stable preference ordering should produce identical rankings of A and B whether it is elicited as a pairwise choice or in terms of dollar valuations. A preference reversal occurs when two modes of elicitation theorized to produce the same implicit rankings actually produce rankings that reverse one another.

One explanation is that when asked to choose, people focus on the risk of getting zero. Gamble A provides a lower risk of getting zero and therefore dominates B by this criterion. When asked to give a dollar valuation, however, people tend to focus on the amounts or magnitudes of payoffs more than the probabilities of their occurrence. Focused on the magnitude of the largest payoff, gamble B's 40 dominates gamble A's 4. These different thought processes—prioritizing risks of getting zero or prioritizing the magnitude of the largest payoff, respectively—might be very reasonable approaches to decision making in particular contexts. Nevertheless, they violate the norms defined by the standard definition of a rational preference ordering (see Brandstätter, Gigerenzer, & Hertwig's [2006] alternative explanation based on their *priority heuristic*).

### Measuring Risk and Time Preferences

Innovative techniques for measuring preferences along a number of dimensions have emerged as an interesting subset of behavioral and experimental economics. Given the widespread reach of expected utility theory in economics in and outside behavioral economics, Eckel and Grossman (2002, 2008) and Holt and Laury (2002) have designed experimental instruments for quantifying the extent to which people are risk averse or risk loving. The Eckel-Grossman instrument has proven useful in capturing interpersonal variation of risky choice in a wide variety of populations, including those with very limited experience interpreting numerical risk measures, thanks to its remarkable simplicity. Together with a parameterized expected utility function, their instrument produces quantifiable ranges for an individual's risk aversion parameter based on a single choice from among six binary gambles, where each gamble has only two possible outcomes, each of which occur with 50% probability. Reconciling possible inconsistencies among risk measures generated by different instruments and exploiting information about subjects whose preferences appear inconsistent is an active area of ongoing research (Berg, Eckel, & Johnson, 2009; Dave, Eckel, Johnson, & Rojas, 2007).

Behavioral economists have also established commonly used techniques for measuring time preferences. In the standard formulation, a person's time preference, or impatience, can be identified as the extent to which he or she trades off larger cash flows arriving in the more distant future in favor of smaller cash flows that arrive earlier. Impatience is frequently quantified in terms of the subject discount rate, although this depends on auxiliary assumptions about the utility function.

## Biased Beliefs

In contrast to the violations of internally consistent logic discussed in the previous section, this section introduces another broad theme in behavioral economics concerning subjective beliefs that are objectively incorrect. The gap between a subjective belief about the probability that an event will occur and the objective probability of its occurrence (assuming an objective probability exists) is referred to as bias. The very notion of biased beliefs depends on how well calibrated subjective perceptions are to external benchmarks (i.e., objective frequencies of occurrence in the world), whereas the rationality assumptions discussed earlier are based solely on internal consistency and make no reference to external normative benchmarks when describing what it means to make a good decision. Studies

of biased beliefs confront surprisingly subtle challenges, first, in measuring people's subjective beliefs and, second, in establishing the existence of proper benchmarks (in the form of objective probabilities) against which subjective beliefs can be compared.

### "All the Kids Are Above Average" Not as Crazy as It Sounds

One sometimes hears people who should know better mistakenly claim that, if most people's beliefs about an attribute of theirs is different from the average value of that attribute, then people's beliefs must be systematically wrong (e.g., nearly everyone reporting that they are better-than-average drivers in terms of safety). In a bell-curved or other symmetric probability distribution, gaps between what most people believe—the modal response—and the average might justifiably be interpreted as evidence of bias. However, in many real-world probability distributions, such as traffic accidents (where a few bad drivers are responsible for most of the accidents) or annual income (where a small number of very high-earning individuals pull average income well above median income), the sample average is surprisingly nonrepresentative of most people.

Consider a society comprising 999 people who have nothing and one person—call him Bill Gates—who owns $1 billion in wealth. The average person in this society is a millionaire, with average wealth = $1,000,000,000/1,000 = $1 million. Nearly everyone in this society is poorer than average. Thus, when the modal belief about how wealthy a person is turns out to be significantly lower than average wealth, it implies no bias in beliefs. Realizing this, researchers attempt to carefully elicit beliefs about medians and other percentiles that pin down the value of some variable $X$ below which a known percentage of the population falls (e.g., Camerer & Hogarth, 1999).

### Bias Implies Existence of Normative Benchmarks

Bias in econometrics is defined as the difference between the expected value of an estimator and the true value of the number(s) being estimated. In econometrics as well as in everyday usage, asserting that there is a "bias" implies having made an unambiguous commitment to what the true or correct value is. The term *bias* is ubiquitous in behavioral economics, and much of its empirical and theoretical work concerns deviations from normative benchmarks that implicitly assert how people ought to behave. Given the observation that people deviate from a benchmark (typically an axiomatic definition of rationality or formal logic), there are at least two distinct reactions to consider.

Most behavioral economists have interpreted observed deviations from the assumptions of neoclassical economics as bias, implicitly asserting that neoclassical assumptions are undisputed statements defining what good, or smart, economic behavior ought to be. According to this view, people who deviate from the neoclassical benchmarks are making mistakes, which is equivalent to saying they are biased. This, in turn, motivates some authors to recommend prescriptive policy changes aimed at de-biasing the choices we make, inducing us to more closely conform to the axioms of economic rationality. Alternative interpretations of observed deviations from neoclassical norms have been put forward by those who question whether the neoclassical model provides sound guidance for how we ought to behave and by those who fear the paternalistic implications of policies aimed at de-biasing choice. One alternative interpretation takes its point of departure from the observation that people who systematically violate neoclassical assumptions are also surviving quite successfully in their respective economic environments—they are going to college, holding down jobs, having children and grandchildren, and so on. If this is the case, then social scientists should abandon neoclassical benchmarks as normative guideposts in favor of more meaningful measures of economic performance, happiness, health, longevity, and new measures of adaptive success that have yet to be proposed.

## Social Preferences

The standard assumption of unbounded self-interest is often described as a hypothesis holding that people only care about their own monetary payoffs and are completely indifferent among allocations of payoffs to different people in a group as long as their own payoff is the same. Challenging this assumption, behavioral economists seeking to study the extent to which people care about the overall allocation of payoffs among participants in a strategic interaction have described their alternative hypothesis as "social preferences." Researchers studying social preferences have sought to remain as close to the standard utility maximization framework as possible, modeling and testing the implications of social preferences by introducing utility functions that depend on other people's monetary payoffs as well as one's own payoff. Two of the most famous experiments in behavioral economics, the dictator game and the ultimatum game, are discussed below. These are formulated as extremely simple two-player games in which the hypothesis that people maximize their own monetary payoff makes a clear prediction. After hundreds of experimental tests in many places and in the presence of different contextual factors, there is widespread consensus that real people's behavior typically violates the hypothesis of own-payoff maximization.

### Dictator Game

In the dictator game, one player is handed a resource endowment, say $10, and then decides how to split or allocate it between himself or herself and the other player. The other player has no choice to make. There is typically no

communication, although both players see the entire structure of the game, which means they know the other's payoffs and how different combinations of actions lead to different payoffs. The game is played anonymously and one time only to avoid motivating players to try appearing "nice" in the expectation of future reciprocation. The player making the decision, referred to as the dictator, can keep all \$10 for himself or herself and give the other player \$0. The dictator can also choose a \$9–\$1 split, an \$8–\$2 split, a \$5–\$5 split, and so on. In versions of the game with unrestricted action spaces, the dictator can keep any amount for himself or herself $K$, $0 \leq K \leq 10$, leaving the other player with a monetary payoff of $10 - K$. The theory that players of games maximize their own monetary payoffs without regard for other people's payoffs makes a clear prediction in the dictator game: Dictators will choose to keep everything, maximizing their own monetary payoff at $K = 10$ and allocating zero to the other player.

However, when real people play this game, the most common choice by dictators is a 50–50 split, even when playing versions of the game with much larger monetary payoffs. This is a clear violation of the hypothesis that people maximize an objective function that depends only on one's own monetary payoffs, and it is typically interpreted as evidence in favor of social preferences. In other words, the gap between the predictions of standard economic theory and the data observed in experiments implies (although this point is open to alternative interpretations) that people care about the monetary payoffs of others. Note that caring about the payoffs of others does not imply altruism or benevolence, so that spiteful preferences that register increased psychic gain based on the deprivation of others is also a form of social preferences.

## Ultimatum Game

In the ultimatum game, a proposer receives an endowment, say \$10, and then makes a proposed allocation. Reusing the symbol $K$ to represent the amount the proposer proposes to keep, the proposed allocation looks similar to the allocation in the dictator game: $K$ for the proposer, $0 \leq K \leq 10$, and $10 - K$ for the other player, sometimes referred to as the responder. Unlike the dictator game, however, the responder has a binary decision to make in the ultimatum game: whether to accept the proposer's proposal or not. If the responder accepts, then payoffs follow the proposal exactly. If the responder declines the proposal, then both players receive zero. Once again, this game is typically played anonymously and only one time to limit the expectation of future reciprocation as a confounding motive when interpreting the results.

As long as the proposal includes any positive payoff for the responder, a responder who maximizes his or her own monetary payoff will choose to accept because even a small amount is better than zero according to the money maximization hypothesis. The subgame perfect equilibrium is for the proposer to offer the smallest positive amount possible to the responder and for the responder to accept. For example, if payoffs are restricted to integer values, the strict subgame perfect equilibrium is uniquely defined by a proposal in which the proposer keeps \$9 and the responder receives \$1, and the responder accepts this proposal even though it is far from an even split.

Contrary to the theoretical prediction of proposers offering a \$9–\$1 split and responders accepting it, the most common proposal in the ultimatum game is an even (or nearly even) 50–50 split. The responder's behavior is especially interesting to students of social preferences because responders typically reject unfair offers even though it leaves them with zero as opposed to a small positive amount. It is this willingness of responders to choose zero by rejecting "unfair" offers of \$9–\$1 and higher that provides one of the most decisive pieces of evidence for social preferences. A common interpretation is that the responder receives more utility from punishing the proposer for having made an unfair proposal than he or she would get by accepting \$1 and leaving the proposer's unfair offer unpunished. This shows that the responder does not make decisions solely on the basis of his or her own payoff but is considering the payoffs of the other player.

## Extending Utility Theory to Incorporate Social Preferences

A well-known approach to modeling social preferences (Fehr & Schmidt, 1999) extends the neoclassical utility function (which typically depends only on one's own monetary payoffs) to include three components: utility from one's own payoff (as one finds in a neoclassical utility function), utility from the positive deviation between one's own payoff and other players' payoffs (i.e., the pleasure of doing better than others), and a third term placing negative weight on negative deviations from other players' payoffs (i.e., displeasure of doing worse than others). Some authors have introduced models that add similar "social preferences" terms to the utility function, for example, placing negative weight on highly unequal allocations, weighted by a parameter referred to as inequality aversion.

Critics of the social preferences program draw on distinct points of view. Binmore and Shaked (2007) argue that the tools of classical and neoclassical economics can easily take social factors into account and need not be set off from neoclassical economics under distinct "social preferences" or "behavioral economics" labels. Although Binmore and Shaked are correct that, in principle, neoclassical utility theory does not preclude other people's payoffs from entering the utility function, the assumption of unbounded self-interest is indeed a key tenet of neoclassical normative theory. The no-externalities assumption (i.e., people care only about their own payoffs, and their actions affect each other only indirectly through market prices) is crucial to the validity of the fundamental welfare theorem, which states that competitive markets are socially efficient. It is difficult to overstate the role that this theoretical result has enjoyed

in guiding public policy toward private versus government provision of services such as health care. Critics rightly point to this theory's reliance on the unrealistic assumptions of no externalities and no information asymmetries.

## Rationality

In the popular press, behavioral economics is often portrayed as a branch of economics that points to systematic irrationality in human populations and in markets in particular. Titles such as *Irrational Exuberance* (Shiller, 2000)*, Predictably Irrational* (Ariely, 2008), and much of Nobel laureate Daniel Kahneman's work documenting deviations from axiomatic definitions of rationality make it easy for nonexperts to associate behavioral economics with irrationality. Indeed, many behavioral economists in their writing, especially when describing their results verbally, use *rational* as a synonym for behavior that conforms to standard economic theory and *irrational* as a catch-all label for behavior that deviates from standard neoclassical assumptions.

One prominent voice in behavioral economics, David Laibson, advocates to aspiring behavioral economists that they avoid describing behavior as "irrational" and avoid the ambiguous term *bounded rationality.* Laibson's admonition is, however, very frequently violated, leading to subtle paradoxes regarding the normative status of the neoclassical model within behavioral economics (Berg, 2003).

In neoclassical economics, a rational preference ordering is defined as any ranking scheme that conforms to the axioms of completeness and transitivity. In choice under uncertainty, many economists—including the seminal contributor Leonard Savage—argue for a strong normative interpretation of what has now become the dominant tool in economics for modeling choice under uncertainty (Starmer, 2005): expected utility theory. The normative interpretation of expected utility theory asserts that choices over probabilistic payoff distributions that can be rationalized as having maximized any expected utility function are rational, and those that admit no such rationalization are irrational.

In choice problems that involve trade-offs over time, choices that can be rationalized as maximizing a time-consistent objective function are frequently referred to as "rational" and those that cannot as "irrational." An important intuitive problem arises in these uses of the term *rationality*—namely, that all parties in a strategic interaction in which more than one party is playing an allegedly "irrational" strategy may be strictly better off than would be the case if each party accepted economic "rationality" as a prescription for action.

Two distinct problems arise. First, there is behavior that conforms to axiomatic rationality but is manifestly bad or undesirable in many people's views. Second, there is behavior that is very reasonable to many people because it achieves a high level of performance, which nevertheless violates axiomatic rationality. Thus, axiomatic rationality is both too strong and too weak. Too strong because it rules out reasonable behavior as irrational, and too weak because it allows for behavior whose consequences for well-being are intuitively bad in many people's views.

Rational preferences impose the requirement of self-consistent choice but typically say nothing about how well choices work in the real world, a distinction that psychologists Hastie and Rasinski (1988) and Hammond (1996) describe as coherence (internally consistent) versus correspondence (well calibrated to the world) norms. Thus, dropping out of college, walking past a pile of cash on the ground, becoming addicted to drugs, or even committing suicide can be rationalized (and regularly have been in the economics literature) as maximizing a rational preference ordering because they can be made to satisfy internal coherence. Economic rationality here imposes nothing more than consistency.

A rational person can walk past $100 lying on the sidewalk, revealing (within the rational choice framework) that his or her disutility of stopping to lean over and pick up the money is greater than the benefit of the money. Rationality requires only that the person is consistent about this ranking, never stopping to pick up money on the sidewalk when the amount is $100 or less. In the other direction, a person who drops out of college and then—without anything else in his or her life circumstances changing significantly—decides to reenroll is regarded as inconsistent and therefore irrational, even though many parents would no doubt regard this as, on the whole, good economic behavior.

In expected utility theory, a decision maker can be completely averse to risk or love risk taking, but not both. The requirement of rationality is that all risky choices are consistent. Thus, someone who always takes risks and perhaps is regarded by many as foolish and imprudent would pass the consistency requirement, conforming squarely with axiomatic rationality based on consistent foolishness. In the other direction, a person who buys health insurance (revealing himself or herself to be risk averse) and also decides to start a new business (revealing himself or herself to be risk loving) cannot easily be rationalized within expected utility theory because of the appearance of inconsistent choices in risky settings.

In time trade-off decision problems, consistency allows for both infinite impatience and infinite patience, but not in the same person. A person who, every payday, throws a party and spends all his or her money and then starves until next payday passes the consistency test and is therefore rational. Consistently strange behavior (e.g., blowing one's paycheck and starving thereafter until the next paycheck) can be rationalized, but inconsistent behavior—even when it seems to reflect a positive step in a person's maturing or taking responsibility for his or her well-being—is labeled irrational because it violates consistency. If the person who previously blew his or her paycheck every payday decides to start saving money for his or her retirement, this would be inconsistent, although very reasonable to most people. These cases illustrate

tension between what most people regard as sensible economic behavior and the surprising simultaneous tightness and looseness of rational choice as it is defined in economics as a criterion for normative evaluation.

## Behavioral Economics: Prospects and Problems

The origins of behavioral economics are many, without clear boundaries or singularly defining moments (Hands, 2007; Heukelom, 2007). And yet, even a cursory look at articles published in economics today versus, say, 1980 reveals a far-reaching behavioral shift. One can cite a number of concrete events as markers of the emergence of behavioral economics onto a broader stage with wide, mainstream appeal. One might imagine that such a list would surely include Herbert Simon's Nobel Prize in 1978. But prior to the 1990s, behavioral work appeared very infrequently in flagship general interest journals of the economics profession. A concise and, of course, incomplete timeline of milestones in the recent rise of behavioral economics would include Richard Thaler's "Anomalies" series, which ran in the *Journal of Economic Perspectives* starting in 1987; hiring patterns at elite business schools and economics departments in the 1990s; frequent popular press accounts of behavioral economics in *The Economist, New York Times,* and *Wall Street Journal* in the past 10 years; and the 2002 Nobel Prize being awarded to experimental economist Vernon Smith and psychologist Daniel Kahneman. The 1994 Nobel Prize was shared by another economist who is an active experimenter and leading voice in game theory and behavioral economics, Reinhardt Selten.

A striking element in the arguments of those who have successfully brought behavioral economics to mainstream economics audiences is the close similarity to Friedman's as-if methodology. In prospect theory, behavioral economics adds new parameters rather than psychological realism to repair and add greater statistical fit to an otherwise neoclassical weighting-and-summing approach to modeling choice under uncertainty. In the social preferences approach, behavioral economics adds parameters weighting decision makers' concern for receiving more, or less, than others do to an otherwise neoclassical utility function. In intertemporal choice, behavioral models of time inconsistency add discounting parameters with nonexponential weighting schemes while hanging onto the assumption of maximization of a time-separable utility function. Frequently billing itself as a new empirical enterprise aimed at uncovering the true preferences of real people, the dominant method in the most widely cited innovations to emerge from behavioral economics can be perhaps better described as filtering observed choices through otherwise neoclassical constrained optimization problems, with augmented utility functions that depend on new functional arguments and parameters.

Behavioral economists' attempts to filter data through more complexly parameterized constrained optimization problems suggest more similarity than difference with respect to neoclassical economics. It will be interesting to see whether moves in the direction of neuroeconomics based on brain imaging data portend more radical methodological shifts (Camerer, Loewenstein, & Prelec, 2005) or rather a strengthening of the core methodological tenets of neoclassical economics (Glimcher, 2003). There is a route not taken, or not yet taken, following Herbert Simon's call to abandon universalizing, context- and content-free characterizations of rational choice in favor of models that explicitly consider the interaction of decision processes and the different environments in which they are used—that is, ecological rationality (Gigerenzer, Todd, & the ABC Research Group, 1999).

Criticisms notwithstanding, a number of new and practical suggestions for designing institutions and intervening to help people change their behavior have emerged from the behavioral economics literature of recent decades. These include plans that encourage greater levels of retirement savings (Benartzi & Thaler, 2004), higher rates of organ donation (Johnson & Goldstein, 2003), controlling the amount of food we eat (Wansink, 2006), and new tools for encouraging greater levels of charitable giving (Shang & Croson, 2009). One recent example of behavioral economics being put into practice is President Obama's tax cut, which is disbursed as a small reduction in monthly tax withholding, thereby increasing workers' monthly take-home pay by a small amount each month instead of arriving in taxpayers' mailboxes as a single check for the year. In the rational choice theory, taxpayers' decisions about how much of the tax cut to spend should not depend significantly on whether one receives, say, 12 paychecks with an extra $50 or a single check for $600 for the entire year. But behavioral economists such as Richard Thaler have advised the Obama administration that, to induce immediate spending (which is what most economists call for in response to a recession), it is important that taxpayers view tax cuts as an increase in income rather than wealth—in other words, that taxpayers put the tax cut proceeds into a "mental account" from which they are relatively more likely to spend (Surowiecki, 2009).

All indications suggest that more empirical findings and theories from behavioral economics will make their way into public policy and private organizations aiming to influence the behavior of workers and consumers. Whether these tools will be regarded as benevolent interventions that make our environments better matched to our cognitive architecture or an Orwellian shift toward psychology-inspired paternalism is currently under debate (Berg & Gigerenzer, 2007; Thaler & Sunstein, 2003). There would seem to be genuine cause for optimism regarding behavioral economists' widely shared goal of improving the predictive accuracy and descriptive realism of economic

models that tie economics more closely to observational data, while undertaking bolder normative analysis using broader sets of criteria that measure how smart, or rational, behavior is.

## References and Further Readings


Ariely, D. (2008). *Predictably irrational.* New York: HarperCollins.

Benartzi, S., & Thaler, R. (2004). Save more tomorrow: Using behavioral economics to increase employee saving. *Journal of Political Economy, 112,* 164–187.

Berg, N. (2003). Normative behavioral economics. *Journal of Socio-Economics, 32,* 411–427.

Berg, N., Eckel, C., & Johnson, C. (2009). *Inconsistency pays? Time-inconsistent subjects and EU violators earn more* (Working paper). Dallas: University of Texas–Dallas.

Berg, N., & Gigerenzer, G. (2007). Psychology implies paternalism? Bounded rationality may reduce the rationale to regulate risk-taking. *Social Choice and Welfare, 28,* 337–359.

Berg, N., & Gigerenzer, G. (In press). "As-if" behavioral economics: Neoclassical economics in disguise? *History of Economic Ideas.*

Binmore, K., & Shaked, A. (2007). *Experimental economics: Science or what?* (ELSE Working Paper 263). London: ESRC Centre for Economic Learning and Social Evolution.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review, 113,* 409–432.

Bruni, L., & Sugden, R. (2007). The road not taken: How psychology was removed from economics, and how it might be brought back. *Economic Journal, 117*(516), 146–173.

Camerer, C. (1999). Behavioral economics: Reunifying psychology and economics. *Proceedings of the National Academy of Sciences of the United States of America, 96,* 10575–10577.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction.* New York: Russell Sage Foundation.

Camerer, C., & Hogarth, R. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19*(1–3), 7–42.

Camerer, C., & Loewenstein, G. (2004). Behavioral economics: Past, present and future. In C. Camerer, G. Loewenstein, & M. Rabin (Eds.), *Advances in behavioral economics.* Princeton, NJ: Princeton University Press.

Camerer, C., Loewenstein, G., & Prelec, D. (2005). *Neuroeconomics: How neuroscience can inform economics. Journal of Economic Literature, 43,* 9–64.

Carmon, Z., & Ariely, D. (2000). Focusing on the forgone: How value can appear so different to buyers and sellers. *Journal of Consumer Research, 27*(3), 360–370.

Carpenter, J., & Seki, E. (2006). Competitive work environments and social preferences: Field experimental evidence from a Japanese fishing community. *Contributions to Economic Analysis & Policy Berkeley Electronic Press,* 5(2), Article 2.

Charness, G., & Grosskopf, B. (2001). Relative payoffs and happiness: An experimental study. *Journal of Economic Behavior and Organization, 45,* 301–328.

Dave, C., Eckel, C., Johnson, C., & Rojas, C. (2007). *Eliciting risk preferences: When is simple better?* (Working paper). Dallas: University of Texas–Dallas.

Eckel, C., & Grossman, P. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior, 23,* 281–295.

Eckel, C., & Grossman, P. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior and Organization, 8*(1), 1–17.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition and cooperation. *Quarterly Journal of Economics, 114,* 817–868.

Friedman, M. (1953). *Essays in positive economics.* Chicago: University of Chicago Press.

Gigerenzer, G. (2008). *Rationality for mortals.* New York: Oxford University Press.

Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: The adaptive toolbox.* Cambridge: MIT Press.

Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Glimcher, P. W. (2003). *Decisions, uncertainty and the brain: The science of neuroeconomics.* Cambridge: MIT Press.

Güth, W. (2008). (Non-) behavioral economics: A programmatic assessment. *Journal of Psychology, 216,* 244–253.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* New York: Oxford University Press.

Hands, D. W. (2007). *Economics, psychology, and the history of consumer choice theory.* Retrieved from http://ssrn.com/abstract=988125

Hanemann, W. M. (1991). Willingness to pay and willingness to accept: How much can they differ? *American Economic Review, 81,* 635–647.

Hastie, R., & Rasinski, K. A. (1988). The concept of accuracy in social judgment. In D. Bar-Tal & A. W. Kruglanski (Eds.), *The social psychology of knowledge* (pp. 193–208). New York: Cambridge University Press.

Heukelom, F. (2007). *Kahneman and Tversky and the origin of behavioral economics* (Tinbergen Institute Discussion Papers 07–003/1). Rotterdam, the Netherlands: Tinbergen Institute.

Holt, C., & Laury, S. (2002). Risk aversion and incentive effects. *American Economic Review, 92,* 1644–1655.

Johnson, E. J., & Goldstein, D. G. (2003). Do defaults save lives? *Science, 302,* 1338–1339.

Jolls, C., & Sunstein, C. R. (2006). Debiasing through law. *Journal of Legal Studies, 35,* 199–241.

Kahneman, D., Knetsch, J., & Thaler, R. (1991). The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives, 5*(1), 193–206.

Katona, G. (1951). *Psychological analysis of economic behavior.* New York: McGraw-Hill.

*Lichtenstein,* S., & *Slovic,* P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89,* 46–55.

Loewenstein, G., & Thaler, R. (1989). Intertemporal choice. *Journal of Economic Perspectives, 3*(4), 181–193.

Plott, C., & Zeiler, K. (2007). Exchange asymmetries incorrectly interpreted as evidence of endowment effect theory and prospect theory? *American Economic Review, 97,* 1449–1466.

Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature, 36,* 11–46.

Shang, J., & Croson, R. (2009). Field experiments in charitable contribution: The impact of social influence on the voluntary provision of public goods. *Economic Journal, 119,* 1–17.

Shiller, R. (2000). *Irrational exuberance.* Princeton, NJ: Princeton University Press.

Shogren, J. F., Shin, S. Y., Hayes, D. J., & Kliebenstein, J. B. (1994). Resolving differences in willingness to pay and willingness to accept. *American Economic Review, 84*(1), 255–270.

Simon, H. (1986). Rationality in psychology and economics. *Journal of Business, 59,* S209–S224.

Slutsky, E. E. (1952). Sulla Teoria del Bilancio del Consonatore. In G. J. Stigler & K. E. Boulding (Eds.), *Readings in price theory* (pp. 27–56). Homewood, IL: Irwin. (Original work published 1915)

Starmer, C. (2004). *Friedman's risky methodology* (Working paper). Nottingham, UK: University of Nottingham.

Starmer, C. (2005). Normative notions in descriptive dialogues. *Journal of Economic Methodology, 12,* 277–289.

Surowiecki, J. (2009, January 26). A smarter stimulus. *The New Yorker.* Available at http://www.newyorker.com

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American Economic Review, 93,* 175–179.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New Haven, CT: Yale University Press.

Thaler, R., & Tversky, A. (1990). Preference reversals. *Journal of Economic Perspectives, 4*(2), 201–211.

Veblen, T. (1994). *The theory of the leisure class.* New York: Dover. (Original work published 1899)

Wansink, B. (2006). *Mindless eating: Why we eat more than we think.* New York: Bantam.

# 85

# Experimental Economics

Seda Ertaç
*Koç University*

Sandra Maximiano
*Purdue University*

Being able to test theories and understand the underlying mechanism behind observed phenomena is crucial for scientific progress in any discipline. Experimentation is an important method of measurement in the natural sciences as well as in social sciences such as psychology, but the use of experiments for gathering economic data is a much more recent endeavor. Economics has long been regarded as a nonexperimental science, which has to rely on observations of economic behavior that occur naturally. Experiments, however, have found their way into the economist's toolkit in the past few decades and are now being employed commonly in mainstream economics research in many diverse subfields such as game theory, industrial organization, labor and development economics, and, more recently, macroeconomics.

The very first experiment in economics is known to have been conducted by Bernoulli on the St. Petersburg's paradox in 1738 (see Kagel & Roth, 1995, for more on the history of experimental economics). However, more formal experimentation started in the 1930s with individual choice experiments and flourished especially with the advent of game theory with the work of von Neumann and Morgenstern (1944) on the theory of decision making and games. Also around that time, the first "market experiments" were run by Chamberlin (1948) to test competitive equilibrium. Gradually, experiments started being used more and more widely in many areas of economics, and the number of experimental research papers published in economics journals has been growing rapidly and is now on par with more "classical" fields such as economic theory. The development of experimental economics as a field has also been parallel to advances in the field of "behavioral economics." Behavioral economics aims at integrating insights obtained from psychology into economic models, frequently uses experiments as a method for collecting data and finding out patterns of behavior that are inconsistent with standard theory, and builds new models that can explain the observed behavior in experiments. The most evident recognition of the importance of experimental and behavioral economics was the 2002 Nobel Prize in economics, which was awarded to Daniel Kahneman and Vernon Smith for their contributions to behavioral and experimental economics, respectively.

But why and when do economists need experiments? One of the main goals of empirical analysis in economics is to understand how different models of economic decision making fare in understanding observed economic behavior and outcomes and therefore test the predictive success of economic theories. However, it is not always possible to conduct proper tests of theories or to measure the effects of different economic policies using naturally occurring data because of at least three reasons. First, naturally occurring data may simply not exist. For example, in testing models of strategic interaction, oftentimes it is important to know what people think or believe about other people's actions, and although actions are observable, beliefs are not. Similarly, reservation wages and workers' outside options are also not observed in naturally occurring data but are important to understand agents' behavior in labor markets. Laboratory experiments, on the other hand, allow us to collect data on these unobservables.

Second, experiments allow *randomization* into treatments of interest, reducing selection bias and giving the researcher the proper counterfactual for causal inference. Consider the following example. Suppose that we are interested in the effects of tournament-type compensation schemes (where one's wage depends on his or her performance relative to others) versus piece-rate compensation schemes (one's wage depends only on his or her own performance) on worker productivity. If we collect real worker productivity data from firms that use each of these two types of incentive schemes and make a direct comparison, we cannot be sure whether any productivity differential we observe comes from the true effect of incentives or from firms or workers' unobservable differences that are correlated with productivity. For example, more able and more ambitious workers may think that they have better prospects in a firm that uses tournament incentive schemes, and this motivation differential will bias the estimates of the "treatment effect" of incentives on productivity. That is, the samples under the two incentive schemes are "selected" and not directly comparable, and this incomparability will blur any type of inference we can make. Although econometricians have been devising methods that could alleviate some of these problems (such as instrumental variables and matching techniques), having direct control over the data-generating process makes inference much simpler. In the context of the incentives example, by assigning workers randomly into two treatments, one where they work under a tournament incentive scheme and one under the piece-rate scheme, it would be possible to measure the "true" effect of incentive schemes on productivity. This is possible because random assignment acts as a control for individual characteristics, considering the fact that if you have a sufficient number of subjects, the two groups will look sufficiently alike along dimensions such as ambition, ability, and so on, which we would like to control. By assigning subjects randomly into "treatments" that differ along the dimension of interest, experiments take care of the selection problem through randomization and can accurately isolate the effect of the focus variable.

Likewise, in a controlled experiment, which variables are exogenous and which are endogenous is clearly known, and this allows the experimenter to make causal inferences about the association between the variables, whereas with natural data, there are usually many factors changing at the same time, making it hard to disentangle the effect of a certain factor on the variable of interest. Another rationale for using experiments, perhaps of more theoretical interest, is that it may be difficult to test theories of one-shot interactions with naturally occurring data since factors such as reputation are oftentimes present because interactions naturally take place mostly in repeated settings. All these advantages suggest that experimentation can be a very valuable tool for gathering data on economic decision making.

While economic experiments have generally used the laboratory as their setting and university students as subjects, "field experiments" have been receiving much attention recently. By testing behavior in a setting that is more "natural" on several dimensions, field experiments provide insight on the "external validity" or generalizability of the results from laboratory experiments and complement lab experiments in improving our understanding of economic phenomena. According to Harrison and List (2004), experiments may differ in the nature of (a) subject pool, (b) information and experience that the subjects bring to the task, (c) the commodity being transacted, (d) task and institutional rules, and (e) the environment that subjects operate in. These five factors are used to determine the field context of an experiment and result in the following taxonomy: (a) *conventional lab experiments,* which use a standard subject pool of university students, abstract framing (e.g., choices are labeled A, B, C, etc., rather than with words that suggest context), and an imposed set of rules; (b) *artifactual field experiments,* which are conventional lab experiments but with a nonstandard subject pool (e.g., actual workers from a firm rather than university students); (c) *framed field experiments,* which are artifactual field experiments but with field context in the task, commodity, information, or environment (e.g., rather than trading a "virtual" commodity on the computer, subjects trade a real commodity in a real field context); and (d) *natural field experiments,* which are framed field experiments in which subjects are not aware that they are in an experiment.

In what follows, we will first provide a discussion of some of the main methodological issues involved in experimental research and then provide a selective account of the main areas of application where experiments have been employed in economics. Because of space issues, we are bound to leave out many important domains and applications, so our treatment here should be taken as an illustrative approach that provides examples of how experiments can contribute to our understanding of economic behavior.

## The Methodology of Experiments

The goal of the experimenter is to create a setting where behavior can be measured accurately in a controlled way. This is usually achieved by keeping constant as many variables as possible and varying others independently as "treatment variables." To achieve as much control as possible, the experimental method in economics is based on "inducing preferences" by using appropriate monetary incentives that are tied to the consequences of decisions made during the experiment. A typical economics laboratory experiment involves recruiting subjects, who are usually university students, and paying them a fixed show-up fee for their participation, plus the amount they earn during the experiment. The amount they earn during the experiment, in turn, is tied to the payoff consequences of the decisions they make. This feature of economics experiments, in fact, is a main characteristic that distinguishes economics

experiments from psychology experiments, where decisions are often hypothetical and not incentivized.[1]

For properly inducing preferences, the experimenter should have control over the preferences of the subjects and should know what the subject is trying to attain. For example, in a game theory experiment, we would like the numbers in the payoff matrix of the game to represent the players' utilities, and we therefore pay subjects according to the payoffs in the matrix. Naturally, however, there may be unobservable components of utility affecting subjects' behavior. For instance, if one runs a 100-period experiment with the same type of choice being repeated every period, subjects might get bored and start acting randomly. Alternatively, some subjects might have preferences over other subjects' monetary earnings, and if they know the payoff distribution among the participants, this might affect their behavior in ways that are unrelated to the hypothesis of interest. To make the effects of these unobserved factors minimal, one should make monetary rewards as strong as possible. Having salient monetary rewards will also ensure that subjects are motivated to think carefully about the decision, especially when the decision task is cognitively demanding. Attaching strong monetary consequences to decisions reduces the occurrence of random decisions and minimizes the errors and outliers that would otherwise be observed more often in the data.

On the basis of these general ideas, Nobel Prize winner Vernon Smith (1982) put forward the following precepts for achieving proper control:

- *Nonsatiation*: This means that individuals should prefer more of the reward medium used in the experiment (usually money) to less of it.
- *Saliency*: This assumption means that the reward medium is suitably associated with the choices in the experiment—for example, if one wants action A to be a better action for the individual than action B, then action A should be associated with a higher monetary payoff than B.
- *Dominance*: This means that the reward structure should dominate other factors associated with participation in the experiment (e.g., boredom). High monetary stakes can help achieve this.
- *Privacy*: Individuals' utility functions may have unobservable components that depend on the utility or payoffs of others. For example, a subject's decisions may depend on how much others are earning, if he or she cares about "fairness" of payoffs across subjects. By withholding information about other subjects' payoffs and giving subjects information about their own payoffs only, this potential issue can be mitigated, and better control of preferences can be achieved.

While inducing preferences is very important for experimental control, there are some cases in which experimenters do not want to induce preferences but are interested in obtaining information on the natural ("homegrown") preferences of subjects. For example, we may be interested in knowing how much an individual values a certain object (e.g., the item for sale in an auction). Likewise, we may be interested in knowing the beliefs of the individual. As mentioned before, the latter can be especially important in testing game-theoretic models, where subjects' beliefs about others' possible actions are important in shaping optimal strategies. Experimental economists have devised incentive-compatible mechanisms to elicit subjects' true valuations for an object (e.g., the maximum amount of money they would be willing to pay to buy the object) and use various techniques to elicit subjects' beliefs truthfully. While discussion of these methods in detail is beyond the scope of this chapter, it is important to note that these methods have allowed economists to obtain crucial information that is not available in the field.

One last principle that is relevant for running a "good" experiment is "design parallelism" (Smith, 1982). This principle refers to the need for laboratory experiments to reflect naturally occurring environments to the extent possible. This is very much related to the concept of "external validity" of the experiment, in other words, how much the experimental decision resembles decisions that individuals face in the "real world," which will affect the extent to which the experimental results can be extrapolated to natural economic environments. Although the external validity of an experiment is important, it should be noted that adding more and more complexity to an experiment to increase external validity could result in loss of control and compromise the "internal validity" of the experiment, making the data useless. One should therefore be careful in choosing an experimental design that can be as realistic as possible while still maintaining proper control and avoiding confounds.

## Applications of Experimental Methods in Economics

### Individual Decision-Making Experiments

The experiments conducted in this area analyze nonstrategic decision making by a single individual, in a context with no interaction between the subjects in the experiment. The goal is to understand the decision-making process of the individual and the motivation behind the observed behavior. The topic is strongly related to the psychology of judgment and illustrates the interdisciplinary aspect of experimental methods quite well. In fact, as we will explain below, this strand of the experimental economics literature has collected quite a large body of observations that are inconsistent with standard economic theory and has been an important part of the research in behavioral economics.

Experiments of individual decision making investigate choice in different contexts: over time, under static uncertainty, under dynamic uncertainty, and so on. A very important set of experiments here concerns decision making

under uncertainty and provides tests of the relevant standard economic theory, which posits that individuals are expected utility maximizers. As mentioned before, this means that individuals maximize the expected value of utility, defined as the sum of utility from different possible outcomes, each multiplied by the respective probability that that outcome occurs. Anomalies, or observations violating expected utility theory in these experiments, have been frequent. For example, a well-known departure from the predictions of expected utility is the Allais paradox. Consider the following example, with two decision tasks (Kahneman & Tversky, 1979):

*Decision 1:*

- Option A: $3,000 for sure
- Option B: $4,000 with 80% chance, $0 with 20% chance

*Decision 2:*

- Option C: $3,000 with 25% chance
- Option D: $4,000 with 20% chance

A vast majority of subjects select Option A in Decision 1 but Option D in Decision 2, which is inconsistent with expected utility theory since the lotteries in Decision 2 are equivalent to Decision 1 (when one divides all the winning probabilities of Decision 1 by 4, Decision 2 is obtained, and this across-the-board reduction in winning probabilities should not affect the choice according to expected utility theory).

Choices in such lottery choice experiments have also pointed to a "reflection effect": While individuals make risk-averse choices when the choices involve gains (e.g., a lottery that pays $100 with 50% chance and $0 with 50% chance vs. a sure gain of $40), they act as if they are risk loving when the choices involve losses (e.g., a lottery that involves a $100 loss with 50% chance and no loss with 50% chance vs. a sure loss of $40). That is, losses and gains are treated differently, which is again inconsistent with expected utility theory.

The laboratory evidence that highlights such anomalies has stimulated the development of alternatives to expected utility theory. A well-known alternative, called "prospect theory," was proposed by Kahneman and Tversky (1979). Prospect theory allows for loss aversion, reference dependence, reflection effect, and probability miscalculations and can explain a significant amount of the anomalous findings in the lottery choice experiments. Through its modeling of reference dependence and loss aversion, prospect theory can also explain a phenomenon called the "endowment effect," which refers to the observation in experiments that there is a discrepancy between individuals' valuations of a good, depending on whether they own it or not. That is, the minimum amount of money that individuals are willing to accept in order to part with a good they own (e.g., a coffee mug that has been given to them in the experiment) is higher than the maximum amount they would be willing to pay for the same good when they do not own it, which is inconsistent with standard theory.

Although many experimental results in individual decision-making experiments point to deviations from the predictions of standard economic theory, there is also some evidence that market experience and large monetary stakes can improve the alignment of observed behavior with standard predictions. For example, John List (2006) finds, in a field experiment involving traders in an actual market for sports memorabilia, that inexperienced traders display a strong endowment effect, but experienced traders who have been engaging in market activities for a long time do not. This might suggest that these anomalies or biases might be less prevalent in actual markets, where individuals self-select into economic roles and gain experience through repeated transactions.

One methodological point to be made here is that the type of individual lottery choice tasks mentioned above can also be used to measure subjects' risk preferences in the laboratory. In many different experiments involving a wide range of economic decisions, it is useful to know the risk preferences of individuals. For example, risk-averse and risk-neutral individuals are predicted to bid differently in auctions, and having information on risk preferences allows economists to obtain better insight into behavior. One such method to measure risk preferences is the "Holt-Laury mechanism" (Holt & Laury, 2002), which involves giving subjects a series of choices between a risky lottery (that has a large spread between the good and bad payoff) and a safe lottery (that has a small spread between the good and the bad payoff), which differ in the likelihood of the good payoff. As the probability of the good payoff increases, the attractiveness of the risky lottery increases. By looking at when the subject switches to the risky lottery as the good payoff probability increases, it is possible to get a measure of the risk aversion of the subject.

## Game Theory Experiments

Most of modern microeconomics research involves models of strategic interaction. Since the work of John von Neumann and Oskar Morgenstern (1944) and later of John Nash, game theory has become the fundamental approach in microeconomics, replacing the Walrasian and Marshallian models of consumption, production, and exchange in which individuals take decisions in response to exogenous prices by models of strategic interaction. The theory of games has been used to analyze strategic behavior in coordination games, public good contribution games, auctions, analysis of oligopolies in industrial organization, and so on. In general, game-theoretic models assume that individuals are rational and have stable preferences over outcomes and correct beliefs as to how choices affect the relative probability of possible outcomes. They are assumed to maximize their own expected payoff given their preferences and beliefs and

take into consideration material and informational constraints that may exist.

There are two main criticisms to game theory. First, individuals may not be able to be as forward looking as game theory predicts. They may not always behave rationally or may not perceive others to be rational. Second, people do not behave "selfishly" in all situations and may not expect others to behave in that way either. Experimental methods help economists explore how important these issues are. In experiments, information and incentives are held constant, which allows us to accurately test how well game-theoretic principles predict behavior and understand in which situations individuals do behave in line with the theory and in which cases they deviate from it.

Many of the experiments on game theory have tested (a) the rationality assumption or the depth of strategic reasoning and (b) the assumption of "selfishness," which posits that individuals' utility is dependent on their own monetary payoffs only. In the following discussion, we will focus on games that have attempted to clearly test these two postulates of standard economic theory. First, we will illustrate the effects of limits on rationality through experiments on limited strategic thinking. Second, we will discuss social preferences experiments that test the assumption of selfishness and exclusively money-maximizing behavior.

## Illustration 1: Why Aren't We All Chess Players? Experiments on Limited Thinking

We will illustrate tests of rationality through a discussion of the "guessing game." An important recent set of experiments that has provided a great way to test the depth of players' reasoning is "guessing games" or "beauty contest games." (This name is based on John Maynard Keynes's likening of the stock market to a newspaper beauty contest where readers' aim is to guess the most beautiful lady, as determined by the population. Keynes noted that in such games, it would be important to guess how others think, how others think others think, and so on.) In the canonical version of the guessing game (Nagel, 1995), a group of players is asked to choose a number from a given range (e.g., [0, 100]). The average of all numbers submitted is taken, and 2/3 (or any fraction smaller than 1) of that average becomes the "target number." The person whose chosen number comes closest to this target number wins a fixed prize. In other words, the goal is to correctly guess 2/3 of the average guess. The prediction of game theory in this game is that all players will submit a guess of zero. In fact, this solution can be achieved by a process called "iterated elimination of dominated strategies," which proceeds as follows: If I am rational, I should realize that since the maximum possible number is 100, the target number can never be more than 66. Therefore, any guess above 66 is a "dominated" action—no matter what others might be doing, I should not submit a guess more than 66. But if I know that everyone is rational, I should also realize that no one will submit a guess above 66, which makes any guess above 44 dominated for me because I know the target number cannot be more than 44. Continuing in this fashion, it is possible to reach the unique equilibrium outcome of everyone guessing zero. Experimental results, on the other hand, show that very few people submit guesses of zero and that there are clusters of observations around points such as 33 and 22. Researchers have provided the following type of model to account for these deviations from the prediction of game theory: Suppose that individuals differ in their "depth of reasoning" and are classified as "Level $k$ thinkers," where $k$ is the number of rounds of iterated reasoning they can engage in. For example, a Level 0 person just guesses randomly, without any strategic thought. A Level 1 person, on the other hand, thinks that everyone else is Level 0 and best responds to that—meaning, a Level 1 person will assume that on average the guess will be 50 and therefore chooses a number that is 2/3 of that, 33. A Level 2 person thinks that everyone else is Level 1 and therefore the average guess will be 33 and takes 2/3 of that to submit his or her guess. The experimental data show that most people are within three levels of thinking. Although this game may seem of theoretical interest, in fact it highlights mechanisms that are important in many economic settings, such as the stock market, where it is important to guess what others think about the value of a stock and how rational they are in their behavior.

## Illustration 2: Do We Care Only About Ourselves? Experiments on Social Preferences

As mentioned before, another important focus of game theory experiments has been testing the assumption of "pure self-interest," which we define as preferences that depend only on our own monetary earnings, independently of what others earn. A very interesting set of results in this literature comes from experiments that highlight "social preferences."

The concept of social preferences refers to the concern (positive or negative) for others' well-being. We distinguish between two types of social preferences. The first type refers to *outcome-oriented social preferences.* Here, individuals care about the distribution of payoffs, and this encompasses pure altruism, inequality aversion, and possibly a concern for efficiency. The second type of social preferences concerns *intention-based reciprocity* (i.e., individuals care about the intentions that drive other players' actions). Here an individual is willing to sacrifice his own material payoffs in order to reciprocate, either rewarding kind (fair) or punishing unkind (unfair) behavior.

Research in experimental economics has helped us understand the nature of attitudes toward social preferences and how these attitudes interact with self-interest.

An important workhorse that is used for studying social preferences is simple games in which subjects have to decide on an allocation of money between themselves and an anonymous other subject. Experiments on social preferences generally study such games, which include the so-called ultimatum game, dictator game, trust game, gift exchange game, prisoner's dilemma, public good game, and modifications to these canonical settings. Subjects make decisions usually for a certain number of periods and are usually rematched with different subjects every period, which enables economists to think about this as a one-shot interaction and abstract from repeated game effects. In the past three decades, a vast number of experiments on social preferences have been conducted either to check the existence of social preferences (in the lab) or to test the robustness of results to different subject pools, stakes, framing, number of players, and other design and procedural variables. More recently, field experiments have been conducted to test the external validity of lab experiments and to obtain insight on the strength of social preferences in the field. In the following, we first describe classical experiments on social preferences and some extensions. We then discuss the related field experiments.

## Ultimatum and Dictator Games

The ultimatum bargaining game has been one of the most widely studied games in the past 25 years. In the basic version of this game (Güth, Schmittberger, & Schwarze, 1982), two subjects bargain over the division of a "pie." The first player (the proposer) has an endowment of money and decides on the amount $X$ to send to an anonymous partner (the responder), who then decides whether to accept it or not. If accepted, both players get their agreed-upon shares. If rejected, both receive nothing. While the rational solution predicts that the proposer should offer the smallest possible share and the responder should accept it, Güth et al. (1982) find that, on average, proposers offer 37% of the pie and that low offers are frequently rejected. Since then, numerous other experiments using the ultimatum game have been conducted, and many possible methodological explanations (stakes, subject pool, nature of the game) for the gap between theory and empirical results have been tested. Results are robust: (a) The modal and median ultimatum offers are usually 40% to 50%, and mean offers range from 30% to 40%; (b) very low offers (0%–10%) and "too fair" offers (51%–100%) are rarely observed; and (c) offers below 20% are rejected half the time.

The equilibrium in the ultimatum game is easy to compute and is exempt from bounded rationality or confusion as possible explanations for the results, which makes the ultimatum game one of the most common experimental designs used for inference about individuals' social preferences. For instance, when a responder rejects a positive offer, this means that his or her utility function has more than only a monetary argument. For example, a rejection of a positive but low offer could reveal a concern for negative reciprocity, as the responder is willing to sacrifice his or her own monetary payoff to punish the proposer's action. When the proposer makes a higher offer, however, it could mean a preference for fairness, a fear of rejection, or both. Further experiments using the so-called dictator game disentangle the two explanations and show that both have some explanatory validity. Forsythe, Horowitz, Savin, and Sefton (1994) conducted a dictator experiment that mainly removes the responder's move from the standard ultimatum game—proposers have the power to decide on the allocation of the pie, and the responder cannot reject. In this game, offers are found to be less generous than in the ultimatum game (the mean allocation is about 20% of the pie), but there is still a significant fraction of people who give positive amounts of money to the other party.

## Trust and Gift Exchange Games

Trust and gift exchange games are both sequential prisoner's dilemma games, and they represent situations where contracts are necessarily incomplete, allowing for a controlled study of the nature and effectiveness of trust and reciprocity in economic interactions. The trust game (Berg, Dickhaut, & McCabe, 1995), which is also called the investment game, considers a situation in which one individual (the investor) transfers to another (the trustee) the power to make a decision that affects the utility of both. More specifically, the investor has an initial endowment $A$ and decides the amount $X$ of $A$ to invest. The money invested is multiplied by $r$ and transferred to the second player, who decides the amount $Y$ of $rX$ to return to the investor. The trustee, therefore, plays a dictator game with an endowment decided by an initial investment made by the recipient. The investor receives $A - X + Y$, and the trustee receives $rX - Y$.

When players are entirely selfish and only interested in maximizing their own monetary payoffs, the second player would never return any positive amount of $rX$. Given that the investor knows that the second player will behave selfishly and he or she will receive nothing in return, Player 1 invests nothing, keeping $A$ for himself or herself. Although it yields a socially inefficient outcome (the total payoffs would be maximized if the investor transferred everything), this "subgame perfect" equilibrium of zero investment holds because (a) players cannot sign binding and irrevocable contracts before the beginning of the game, and the trustee cannot be punished for not sending a positive amount back to the investor, and (b) the game is played only once or with different partners, so there is no role for reputation formation and strategic behavior on the part of the trustee.

In Berg et al. (1995), each player was matched only once with another player, and subjects' anonymity among themselves as well as anonymity from the experimenter was guaranteed. Parameters used in their experiment consisted of $10 for investor's endowment, and any amount passed to the trustee was tripled by the experimenter. The results

deviate substantially from standard predictions, supporting both the hypothesis of pure trust on the investors' side and the hypothesis of trustworthiness on the trustees' side. With respect to investors' behavior, the average amount sent was 5.2, with only 2 of 320 investors sending zero, but the amount sent varied substantially across subjects. The trustees returned on average 4.6 (about 1/3 of the tripled amount), and the amount repaid was highly heterogeneous also, with 50% of trustees returning less than $1. The trust game has been often replicated with different subjects and with a fair amount of variation in experimental procedures, either to test for the robustness of the Berg et al. results or to infer about the effectiveness and nature of trust and trustworthiness in different settings with incomplete contracts.

While the first player's behavior in the trust game allows us to infer whether the investor trusts his or her experimental partner, it is less obvious what we can infer about the trustworthiness of the second player. If the trustee chooses to send a positive amount back, it can signal either altruism (a pure concern for the investor's payoffs) or positive reciprocity (the desire to be kind to someone who was kind). Comparing the amount sent back in the trust game with dictator allocations of comparable size endowments, James Cox (2004) found support for both altruism and intention-based positive reciprocity. Further experiments have been conducted on the importance of intentions using a similar game to the trust game, called the "gift exchange" game.

As in the trust game, the gift exchange game (Fehr, Kirchsteiger, & Riedl, 1993) represents a situation where contracts are incomplete. It represents the interaction between an employer and a worker, and it was designed to test efficiency wage theory (Akerlof, 1982), according to which firms will offer higher than market clearing wages, expecting that workers will work harder in return. Workers then compare the wage received with a norm they consider fair and choose whether to increase their effort or not, resulting in a positive wage effort relationship.

In the most basic version of the gift exchange game, the employer first decides on an unconditional wage transfer. After observing the wage that he or she will earn, the worker subsequently decides how much effort to supply. Effort increases the profits to the employer, but is also (increasingly) costly to the worker. As an example, suppose that firms earn $(q - w)e$ and workers earn $w - c(e)$, where $c(e)$ is a convex effort cost function over effort levels that range from 0.1 to 1.0. Fehr et al. (1993) implement this experiment in a labor market where there is an excess supply of workers (eight workers for six employers). The market is organized as a one-sided posted offer, in which workers accept or reject offers in a random order. Therefore, if the worker is entirely selfish, he or she will not supply any effort at all, irrespective of the actual wage offered. Anticipating this entirely flat wage effort schedule, the employer offers the lowest possible wage that satisfies the worker's participation constraint.

Experimental findings sharply contrast these theoretical predictions. Workers are typically willing to supply more effort when a higher wage is offered, yielding a significantly positive correlation between wages and effort, which can be interpreted as positively reciprocal behavior. Results in these games have been highly replicated and appear to be robust and found in various versions of the standard gift exchange game.

In general, the robustness of laboratory results on social preferences has led to the extrapolation of prosocial behavior to a large number of real-world situations. As pointed out by Levitt and List (2007), however, behavior in the laboratory may differ from that observed in real economic environments, depending on the presence of moral and ethical considerations, the nature and extent of scrutiny of one's actions by others, the stakes of the game, self-selection of the individuals making the decisions, and the context in which the decision is embedded. In general, real-world situations tend to be far more complex than those created in simple laboratory games—for example, workers in reality work in large firms with many other workers, at different hierarchical levels and within a complex payment structure. The possible dependence of fairness and reciprocity considerations on the features of such complex environments could invalidate the direct generalizability of laboratory experimental results to natural markets.

More recently, field experiments have been conducted to address such issues of generalizability and to explore the external validity of laboratory experiments on social preferences. One important example of this endeavor is field experiments on gift exchange. Findings from these field experiments appear to be mixed in terms of their support for social preferences. To capture gift exchange in the field, List (2006) conducted a series of field experiments in a sports card fair, where buyers and sellers of sports cards interacted. The buyer asked the seller for a card of a certain (unverifiable) quality and offered either a high or a low price, and the seller chose the quality of card to provide after seeing this offer. When there was no possibility of reputation building, List did not find reciprocal behavior by sellers: The average card quality provided by nonlocal dealers (who had no incentives to build reputation) was the same for both the buyers who offered a high price and the buyers who offered a low price. On the other hand, in a charity donation context, Falk (2007) found that a small gift included in solicitation mails significantly increases charitable donations.

Another issue that has been explored by field experiments is the time dimension of social behavior, which is a dimension of labor interactions largely ignored in laboratory experiments. Gneezy and List (2006) tested the gift exchange hypothesis in two actual spot labor markets, a library data entry job and a door-to-door fundraising job, and investigated the effect of the duration of the task. Both experiments consisted of a control group that was paid according to the promised wage and a treatment group that was told after recruitment that the wage was raised to a higher level than promised. The authors found that paying

more than market-clearing wages had a positive effect on the effort exerted but that the effect was short-lived in both tasks. In particular, treated workers logged more books and raised more money in the initial hours of the tasks, but no significant differences were observed in later hours. Placing the labor relation within a firm instead of a spot labor market, however, Bellemare and Shearer (2009) got different results. They conducted a field gift exchange experiment within a tree-planting firm where workers received a surprise bonus. Here, reciprocity seems to play a role, as workers' average daily productivity increased by 10%. Moreover, workers' reciprocal behavior persisted several days after the gift. The mixed results obtained from gift exchange field experiments suggest that reciprocal behavior may be less important in spot markets and more relevant in economic settings where norms of giving apply such as employment relationships and charitable giving.

## Market Experiments

Market experiments, for which Vernon Smith received his Nobel Prize, have been very important in the development of experimental economics since these marked one of the first rigorous uses of experiments to understand economic behavior. One main motivation for running market experiments is to test the predictions of competitive equilibrium (the equilibrium price realizes where the demand curve intersects the supply curve), the equilibration process and speed, and market efficiency (the question of whether all potential gains from trade can actually be exploited). Another motivation is to study different "market institutions," which are specifications of the rules of trading. For example, a retail market could be organized as a "posted-offer" market where one side of the market (e.g., sellers) posts prices. An alternative market institution is a "double-auction mechanism," where both buyers and sellers could post bids and asks, and all participants can see the highest outstanding bid and the lowest outstanding ask. The design of a typical market experiment provides a direct illustration of induced values. In these experiments, subjects are assigned the roles of buyers and sellers and trade a virtual object. Buyers are assigned "valuations" for the object, whereas sellers are assigned "costs" of selling the object. A buyer's profit, if he or she buys, is his or her valuation minus the transactions price, and a seller's profit, if he or she sells, is the price minus his or her cost. The major finding from market experiments is that competitive equilibrium works (i.e., the trade price realizes at the intersection of the induced demand and supply curves) even with only a few buyers and sellers.

Another sales mechanism that has been analyzed extensively using experiments is auctions. These experiments test bidding behavior under different auction formats such as first-price, second-price, English, and Dutch auctions. The usual experimental design involves assignment of valuations to bidders randomly and the bidders submitting bids to win an experimental object in the auction. The profit of the winner is determined according to the auction format—for example, when the auction is a first-price auction, the winner gets an amount equal to his or her valuation minus his or her bid. The experimental auction literature to date has found some robust results regarding bidding behavior, such as bidders overbidding with respect to the risk-neutral Nash equilibrium prediction in first-price auctions, and different models of behavior have been considered to explain these results, such as risk-averse bidding, or bidders who obtain additional utility when they win ("joy of winning," see Kagel and Levin [2008] for a review). Auction experiments also highlight the feedback between theory and experiments: Experiments test existing theories of bidding, and the results obtained could suggest new models that can more appropriately predict behavior in auctions.

It should also be noted that experiments are also being conducted in fields such as development economics and macroeconomics. Randomized field experiments in development economics have been very useful to test the effectiveness of different policies, such as the use of monetary incentives to increase school performance and attendance (for a review, see Duflo, 2006). Likewise, experimental macroeconomics research has yielded some important insights about the process of equilibration, equilibrium selection, and coordination problems (for a review, see Duffy, 2008).

Years of experimental economics research have also uncovered some differences in behavior driven by individual characteristics. One of these factors is gender. Although men and women behave similarly in many settings, there are some contexts that produce interesting and gender differences. One important context is decision making under risk: Women are found in several studies to be more risk averse than men (Croson & Gneezy, 2008). Likewise, there is some evidence that women are more averse to competitive situations (such as working under a tournament incentive scheme as opposed to an incentive scheme that only depends on one's own performance) than men (Niederle & Vesterlund, 2007), but whether these differences come from biological (nature explanation) or social factors (nurture explanation) is not determined conclusively yet, and this has been an active area of research in recent years (Gneezy, Leonard, & List, 2009).

Another exciting field of research that adds one other layer to experimental methods in economics is neuroeconomics. Neuroeconomic research allows economists to get at the neural basis of economic decision making, which is especially helpful when there are competing models of behavior that make different assumptions about the decision-making process. In this interdisciplinary field that is rapidly growing, subjects' brain activity is measured while they make decisions during an experiment. The predominant method of measurement is functional magnetic resonance imaging (fMRI), although positron emission tomography (PET) and electro-encephalography (EEG) have also been employed. The regions of the brain that are "activated" while making a certain choice can provide economists with direct data that can help them understand the motivations behind the observed behavior. An example of this type of result that has received much interest is one that establishes the source of "non-selfish" preferences in the brain. In an ultimatum game

study by Sanfey, Rilling, Aronson, Nystrom, and Cohen (2003), subjects presented with unfair offers have been found to have different activation patterns in their brain, depending on whether the offer is coming from another human or a computer. In particular, brain areas associated with negative emotion such as anger "light up" when one faces an unfair offer, and this activation correlates with the subsequent decision to reject the offer. Neuroeconomics research can therefore shed light onto what kinds of decision processes are involved in economic choice and can help economists build models that have assumptions that have a "basis" in the brain.

## Conclusion

Given that naturally occurring data are not always sufficient for measuring economic behavior or inference about a variable of interest, controlled experiments provide an invaluable tool for the economist for gathering data. Experimental data can be used for testing existing economic theories, and the results can direct the creation of new ones. For example, many theories of social preferences today have been built to explain experimental findings, and these new models are being increasingly widely applied in areas such as organizational and personnel economics. With the use of field experiments in conjunction with laboratory experiments, it is possible to test economic behavior in a realistic way, obtain insights that can tell us something about the economic behavior of real agents interacting in natural settings, and evaluate the potential effectiveness of different economic policies. We therefore believe that experimentation is likely to continue to establish itself as a standard method of empirical economic research in the years to come.

## Note

1. It should be noted, as an aside, that although some of the topics studied are common (e.g., decision making), the norms of experimental methodology in the two disciplines can be quite different. Another such major difference of economics experiments from psychology experiments is a norm against deception—economists think that having been given misleading information would lead subjects to mistrust and try to second-guess the experimenter in future experiments, which would result in loss of control and contaminate the data in the long run.

## References and Further Readings


Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics, 97,* 543–569.

Bellemare, C., & Shearer, B. (2009). Gift giving and worker productivity: Evidence from a firm-level experiment. *Games and Economic Behavior, 67,* 233–244.

Berg, J., Dickhaut J., & McGabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10,* 122–142.

Camerer, C. (2002). *Behavioral game theory: Experiments in strategic interaction.* Princeton, NJ: Princeton University Press.

Chamberlin, E. H. (1948). An experimental imperfect market. *Journal of Political Economy, 56*(2), 95–108.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior, 46,* 260–281.

Croson, R., & Gneezy, U. (2008). Gender differences in preferences. *Journal of Economic Literature, 47*(2), 1–27.

Duffy, J. (2008). Experimental macroeconomics. In S. N. Durlauf & L. E. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed.). New York: Palgrave Macmillan.

Duflo, E. (2006). Field experiments in development economics. In R. Blundell, W. K. Newey, & T. Persson (Eds.), *Advances in economics and econometrics* (pp. 322–348). New York: Cambridge University Press.

Falk, A. (2007). Gift exchange in the field. *Econometrica, 75,* 1501–1511.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics, 108,* 437–459.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior, 6,* 347–369.

Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica, 77,* 1637–1664.

Gneezy, U., & List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica, 74,* 1365–1384.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3,* 367–388.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42,* 1009–1055.

Holt, C. (2006). *Markets, games and strategic behavior.* Reading, MA: Addison-Wesley.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review, 92,* 1644–1655.

Kagel, J. H., & Lewin, D. (2008). Auctions (experiments). In S. N. Durlauf & L. E. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed.). New York: Palgrave Macmillan.

Kagel, J. H., & Roth, A. E. (1995). *The handbook of experimental economics.* Princeton, NJ: Princeton University Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica, 47,* 313–327.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21*(2), 153–174.

List, J. A. (2006). The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy, 114,* 1–37.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review, 85,* 1313–1326.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics, 122,* 1067–1101.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300,* 1755–1758.

Smith, V. (1982). Microeconomic systems as an experimental science. *American Economic Review, 72,* 923–955.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

# 86

# Complexity and Economics

Troy L. Tassier
*Fordham University*

Complex systems research is a growing field in economics as well as other social and natural sciences. Complex systems research aims to understand underlying phenomena that regularly occur across various complex systems, whether those systems occur in physics, chemistry, biology, social sciences, or any other discipline. Thus, throughout the field, one finds models and methodologies being shared across disciplines. For instance, one finds statistical mechanics methods from physics applied to the study of infectious diseases in epidemiology. The study of economics from a complex systems or complexity perspective is closely tied to the study of complex systems in general.

Defining what is meant by a complex system or complexity can be a difficult task. When one looks up the root word *complex* in a dictionary, one will read something similar to "made up of complicated interrelated parts" or "involved and intricate." These definitions give one a start in defining how the word *complexity* is used in the sciences, but one needs a bit more. A complex system is made of interacting parts (usually many), but those parts do not always need to be complicated. Sometimes, seemingly simple systems produce complex behavior. Further, an understanding of each of the constituent parts individually does not lead to an understanding of the entire system. Thus, complex systems research in economics and elsewhere often constitutes a holistic approach to understanding economic systems. It encompasses understanding not only how individual constituent parts (such as individuals or firms) operate or behave but also how those operations or behaviors aggregate to create a system (such as a market outcome or dynamic time series).

## What Makes a System Complex?

There are several common features of systems that are associated with complex behavior. These are diversity, structured interactions, bounded rationality, adaption and learning, dynamic systems, and lack of centralized authority. These features are inherently associated with many of the systems that economists often study. It is not the case that all complex systems contain all of the elements listed above, but most contain several of these features. In the next section, this chapter describes these common features of a complex system and gives relevant examples from economics.

### Diversity or Heterogeneity

Complex systems frequently are composed of diverse or heterogeneous elements. Elements of a system may be diverse simply because they perform different functions in systems, individuals, firms, and governments, for example. But even agents or objects within a given group or class in a complex system tend to behave, learn, or organize in a multitude of ways. In economics, at the lowest level, the constituent parts are individuals. Individuals in economic systems differ in so many ways it is difficult to count or list them all. Sometimes, this variety is based on characteristics that an agent is born with, such as race, ethnicity, gender, age, or religion. Other characteristics are primarily choices of the agent. Some of these include training and education, residential location, or specialization in a profession, to name a few. Of course, some of these items are interlinked. For instance, one may be born into a Catholic family but convert to another religion. Further,

some of the choice characteristics are constrained choices, where the constraints may vary across individuals. For instance, one's opportunity to acquire education and human capital are constrained by one's ability to pay college tuition or by the educational opportunities provided by one's caregivers in the home, and one's choice of residential location is constrained by the income and wealth that one attains or inherits. Finally, in many economic models, agents are assumed to have homogeneous preferences or tastes. Very few economic models consider that individuals may have different preferences or objectives, yet if one asked each individual in a group of 1,000 to name a favorite restaurant, flavor of ice cream, time to awaken in the morning, and number of hours to relax in a day, one would probably get 1,000 different answers. In summary, there is a great number of ways in which individual economic agents differ, and complex systems models in economics often embrace this diversity.

## Structurally Interacting Agents or Parts

In many complex systems models, there exists a specified structure on which interactions occur. Sometimes, this is based on geography; other times, the interactions are a function of some other structural constraint such as the neurological connections in the human body. In sum, there is some network architecture central to the interactions of agents.

Although this is currently changing, most traditional economic models assume that agents interact without attention to the details of the interaction. In some cases, agents interact only through a market mechanism like a traditional Walrasian auctioneer. In other instances, agents are supposed to interact through random meetings, as though in a so-called Walrasian soup. For instance, most labor market search models have random meetings between potential employers and employees. Agents then optimize their choice to accept employment if offered a job, based on their expected waiting time for other (preferably better) job opportunities to randomly arrive at or above a reservation wage. However, it is common for individuals to learn of job opportunities through family and friends. In fact, around 50% of jobs are found in this manner (Granovetter, 1995). Further, individuals base many of their decisions on information gained from friends, such as recommendations on which products or brands of products to buy, what restaurants to try when visiting a new city, or what theater performances or movies to attend.

There are multiple ways that one can think of this idea of network architecture and the influence of social contacts playing a role in economic outcomes. The first is probably the least controversial: Network structures can act as constraints on the decisions of agents. If information travels through social contacts, then the contacts of an agent help to determine what information that agent holds, whether it is about jobs, products, or another item of economic interest. Second, one may also view the social contacts of an agent in helping to determine one's behavior through a traditional externality perspective. For instance, if all of one's friends own Apple computer products, it is more beneficial to own an Apple than if all of one's friends own PC-based products because of direct reasons such as the (legal or illegal) sharing of software or because of less direct reasons such as troubleshooting when operational problems arise. Third, one can be influenced just by the actions of friends through conformity effects. Rationally, it may be advantageous to attend movies that one's peers attend just so that one can fit in and engage in conversations about these movies. Perhaps less rationally, one can imagine a lemminglike scenario where one attends movies just because his or her friends attended these movies. In any case, the interactions with one's social contacts help determine behavior and decision making either through a traditional constraint-based approach or through less traditional conformity effects.

These social contacts may be exogenously given (e.g., family) or endogenous; they may be a constrained choice of the agent, such as the friendships formed at school. One gets to choose one's friends from the set of other agents that one meets. But this choice is constrained by the opportunities that one has (geography) and, in some cases, by the willingness of other agents to reciprocate the interaction (friendships). In other cases, reciprocation is not needed. For instance, one can pass on some infectious diseases, such as influenza, without asking permission of the recipient agent.

Finally, since each individual has a unique set of friends and family, each person has a unique set of social contacts. Thus, social contacts are yet another way in which agents are diverse.

## Bounded Rationality

Again, as in the discussion of structured interaction, traditional economic models often make simplifying assumptions that are unrealistic. This occurs again when one considers the rationality of agents. Many economic models assume that the agents that populate them can compute the answers to very complicated problems almost instantly. Even before the interest by economists in complex systems-style economics, this überrationality was already being challenged by numerous economists, resulting in the development of a boundedly rational economics model (Rubinstein, 1998). This literature took several directions, and a few of them are discussed here.

### *Limited Cognition*

Most individuals do not often perform complicated calculations like inverting a large matrix in their heads in a matter of seconds. Thus, some boundedly rational economics models simply assume that economic agents do not have the capacity to easily perform some calculations. There have been a variety of ways in which to implement nonrational agents. Perhaps one of the most notable is the "satisficing" approach of Herbert Simon (1996). Here, agents accept a solution to a problem or an alternative that

yields an acceptable but perhaps not optimal solution or alternative.

### *Rational to Be Irrational*

In some cases, the primary constraints on optimization and rationality concern the ability to gather information as opposed to the ability to process this information. For instance, if one wants to find the best price on a common consumer good, say a specific make and model of shoes, in a large city, one could potentially check every retail establishment that sells shoes. But the time cost required to do so may make it irrational to actually complete this comprehensive search for the optimal price. Thus, consumers may find it optimal to simply accept a reasonable price once found. Similarly, in labor economics, many models assume that agents search for jobs and calculate the optimal search behavior, given knowledge about the distribution of offers that exist; they ask themselves if they should accept the jobs offered or wait for better ones, given that they know the distribution of jobs that are available. This line of research still assumes that agents act rationally and also that agents solve a rather complicated optimization problem.

### *Limited Information Due to Network Constraints*

Recall the previous discussion about the transfer of information across social networks. If the majority of information needed to solve a problem or to maximize utility or profits must be obtained from an information source that depends on some contact structure, then agents may optimize, but they do so with potential information constraints. The cognitive abilities of agents may be very powerful, but the agents can act only on the information that is available to them through the given interaction structure. In this scenario, agents may act rationally, given their information, but the actions may appear irrational to outsiders because of the limited information held by an agent. A simple example of this process is contained in the information cascades literature. Here, agents must choose between two similar goods, where one good is superior to the other. Agents receive a private noisy but informative signal about the quality of two goods and observe the choices of other agents choosing prior to them. Through the other agents' actions and the agent's own private signal, the agent rationally calculates the likelihood of each good being the superior one and chooses that good. Even though agents act rationally, it is possible that the agents may coordinate on a bad equilibrium where the agents all choose the inferior good. (See Holt, 2007, for simple examples.)

### *Adaption or Learning*

Central to the idea of bounded rationality is the fact that agents must face some constraint on their ability to optimize. The constraint may be limited cognitive abilities, limited information, or limited time. All of these items force an agent to act in a way that does not guarantee optimization. In some of the cases described, agents simply act rationally given the constraint. But since many if not most economic models include a dimension of time, an alternative and increasingly popular approach is to allow agents to learn over time. The learning often takes on two different dimensions that this chapter calls *experiential* and *imitative* learning. With experiential learning, the agent uses his or her past experiences to try to improve on economic outcomes. This approach may include some type of trial-and-error learning where agents apply a heuristic to a given problem and view the results. Then when faced with the same problem, or a similar problem, the agent adjusts his or her behavior to attempt to reach a better outcome. Or it may proceed according to a more traditional economics approach where agents use a rational cognitive model where an optimal decision is made according to the available information at a given time. Then as more information is revealed, the agent reoptimizes according to the new information. With imitative learning, agents use the experiences of others around them to try to improve their outcomes. For instance, one might copy the strategy of a neighbor who has fared well in a labor market in one's own search for a job. Imitative learning often takes on many dimensions. For instance, one can copy the actions of neighbors or the strategies of neighbors. This distinction must be considered carefully especially in light of the fact that actions are often more observable than strategies.

There are several common methods for incorporating either form of learning. Most involve some type of error or mutation process that allows agents to improve. For instance, suppose that one observes the outcomes of a selection of agents in a population (an agent's neighbors). One method of learning would be for the agent to simply replicate the agent he or she observes who has the best outcome. Another would be for the agent to replicate the best agent in most cases but sometimes make a mistake and, when doing so, replicate another randomly chosen agent. Or in another method yet, an agent could choose another agent to replicate as a function of the other agent's performance; better performing agents are more likely to be replicated than poorly performing agents, but all agents have some positive likelihood of being replicated. The ability of agents to make mistakes often allows the population and the individual agent to improve on their outcomes in the long run. For instance, suppose that every agent can observe the outcome of every other agent in a population. If every agent copies the best agent, then all agents will have the same strategy in the next time period, and no additional learning can result. This is fine if the best agent was acting optimally, but if the agent was not acting optimally, then the population will never act optimally. Thus, one can get stuck with a suboptimal strategy. Replication of some subperforming agents can maintain diversity, which may allow the population to reach a better long-term outcome. Another component common in many learning models is the ability of agents to simply make mistakes or errors. Similar to not always replicating the best

agents, errors can allow for continued diversity in a population and the ability to more fully explore the set of possible solutions to a problem or game. It is common for learning models to incorporate both of these elements. For instance, a genetic algorithm mimics the reproduction found in nature (Holland, 1995). Better performing agents are more likely to be reproduced. But when an agent does reproduce, it does not produce an exact copy of himself or herself; mutations to strategies occur, and sometimes, strategies of agents are combined (as in genes of an offspring being a combination of parents' genes). Finally, there is also a literature that discusses things like optimal rates of learning and optimal rates of errors. Further, it is sometimes best for rates of error to change across time. One may want a high rate of error early on in order to explore and cover a large range of the possible solution space, but once so-called good solution regions are identified, it may be best to begin to limit errors so that these good regions can be better and more fully explored. (See the discussion of simulated annealing in Miller & Page, 2007.) Thus, there can be a balance between the amount of exploration and exploitation in problem solving (March, 1991).

### Dynamic, Complex Adaptive System

Models in complex systems are almost always dynamic. The preceding paragraph listed various ways in which agents learn. Learning is inherently a dynamic phenomenon. But dynamic properties of complex systems are not limited to this area. It is common that agents in a complex system model are changing in some way. Sometimes this change includes learning. But it may also include change in the form of new interactions for the agents, revelation of new strategies, or the creation of new types of agents, firms, or institutions in an economic or social system. Thus, not only are individual agents often evolving but the systems that guide the agents also are changing or evolving. Of course, this produces feedbacks between the system and the agents that make up the system. A strategy or action that does well today may not be the best strategy tomorrow or next year.

As an example, consider Brian Arthur's (1994) "El Farol Bar Problem." In the problem, individual agents in a population of fixed size must decide whether to attend an event (an Irish music night at a local bar, in the original example). If more than $x$% of agents attend the event, each agent that attends receives less utility than if he or she stayed home; the event is too crowded. But if less than or equal to $x$% of agents attend, then each agent attending receives more utility than if he or she had stayed home. The dilemma here is that if all agents use the same strategy, then everyone attends or no one does. More importantly for this discussion, the best choice for each agent depends on the choices of all other agents. Thus, an agent's best strategy today may not be a successful strategy tomorrow if other agents learn and adapt to the system. For instance, suppose that agents rely on their friends to report attendance at the event to them in order to project attendance in the following week. If $x = 75$ and 50% of agents attend this week and honestly tell their friends that they had a great time, then one might expect to get more than 75% attending next week. This may lead agents to develop more sophisticated strategies that may include misinforming other agents. Further, the organizers of the event may also have an incentive to report attendance figures that may or may not be accurate in order to maximize attendance according to a profit function. The important thing to note is that even some simple scenarios or games can easily lead to complex behavior.

### Lack of Top-Down Administration

In most examples of complex systems, there does not exist a central authority that is responsible for overseeing and coordinating activities of the various agents in the system. Thus, outcomes in the system occur as a result of ex ante uncoordinated actions. Coordination may occur in the system, but coordination does not occur as a result of an exogenously specified central authority. (It is possible, though, that such an authority may emerge from the activities of the system.) Thus, complex systems modeling is sometimes referred to as social science from the bottom up (Epstein & Axtell, 1996).

## Economic Outcomes From a Complex System Perspective

One of the hallmark features of most complex systems is that one cannot understand the whole by independently understanding the sum of the parts. For instance, one can understand the incentives of buyers and sellers participating in a market as well as the rules or laws that define a market but still not fully understand the aggregate behavior of the market. In traditional economics, one might focus on something like a price as a market outcome and take this observation as an indicator of market performance or behavior. The study of complex systems embraces a larger goal of also understanding how the market emerges, how relationships between participants may form and fail, and how institutions, laws, consumer strategies, and firm organization change over time.

One of the reasons that complex systems are difficult to understand is because the interactions between system components tend to create complicated feedbacks and nonlinear relationships between various parts of the system. Another hallmark feature of many complex systems is the existence of multiple feedback relationships in the system. Feedback can be positive or negative. Positive feedback exists in a system if a change in variable $x$ causes the system to respond to the change by creating further change in $x$ in the same direction. Negative feedback exists in a system if a change in variable $x$ causes the system to respond to the change by reversing the direction of the change in $x$. As an example of positive feedback, consider the juvenile crime rate in a neighborhood. Suppose that this crime rate is affected by the number of businesses in the neighborhood. More business

activity leads to more jobs for young people, which lowers the crime rate. But a lack of businesses leads to fewer job opportunities and more crime. Further suppose that businesses are adversely affected by high crime rates. Now suppose that an exogenous change occurs in the system and several new businesses open in a neighborhood. By the relationships described, crime rates would decrease; this would lead to even more businesses entering the neighborhood and a further reduction in crime rates. As an example of negative feedback in the same example, suppose that the decrease in the crime rate is met by a decrease in enforcement of laws. This lax enforcement might then lead to an increase in the crime rate.

Of particular importance, positive feedbacks can lead to there being multiple equilibria in a system. As another example of positive feedback, suppose that there are two competing operating systems for a computer, X and Y. Further, suppose that the value to an agent of using a given operating system increases in the number of other agents who use the same system. Thus, if a large percentage of consumers use System X, the value of an agent's using System X is higher than if a small percentage were using System X. One can think of the same scenario with Operating System Y coming to dominate the market. Thus, an equilibrium could be reached where there are a large percentage of Operating System X users and a small percentage of Operating System Y users. Or one could have the opposite scenario, with a large percentage of System Y users and a small percentage of System X users.

## Understanding Complex Systems: Theory and Policy

The possibility of multiple equilibria in a complex system makes the issue of equilibrium selection even more important. For instance, in the simple supply and demand model taught in introductory economics courses, there is only one equilibrium and thus one prediction for the outcome of that model. But if a system has a multitude of equilibria, how does one make predictions? Further, how does one understand the process of selecting and attaining a specific equilibrium?

To begin answering these questions, one needs to consider that not all equilibria are created equal. An equilibrium that is associated with positive feedback is inherently unstable, while an equilibrium that has negative feedback surrounding it is stable. As an example, consider the following equilibrium: a pencil perfectly balanced on its end. This is an equilibrium for the pencil: As long as no one changes the conditions around the pencil by perhaps blowing on it or shaking the table on which it is balanced, the pencil will stay balanced as it is. But if the pencil tips just a bit, the positive feedback introduced by gravity will lead the pencil to tip a little further and eventually fall over. Positive feedback in any one direction away from an equilibrium can result in forces that drive a system away from the equilibrium. On the other hand, negative feedback is associated with stability. Consider the simple supply and demand model of an introductory economics course. In this model, prices act as negative feedback. If the price deviates from equilibrium, perhaps by dipping too low, the shortage created in the market acts to push prices upward and back to equilibrium. On the other hand, if prices increase, the surplus created in the market will pull prices back down toward the equilibrium. So if prices deviate in any direction away from equilibrium, the negative feedback in the system acts to restore the system to the equilibrium.

One way to predict which of the many equilibria in a system will occur is to consider whether an equilibrium is stable or unstable. Like the pencil balanced on its end, any unstable equilibrium requires very specific conditions to occur. If any of these conditions deviate slightly, the system leaves that equilibrium. As an example, throw a pencil in the air and let it land on the desk 100 times; how many times does it land balanced perfectly on its point? Unstable equilibria are almost never observed. Thus, when predicting which equilibrium will occur in a system, one can rule out any unstable equilibria as good predictors for the system.

However, one may still be left with many stable equilibria in a system. Which equilibria will be attained is another focal point for complex systems research, and many angles have been taken in addressing this issue. There are formal theoretical interests such as measuring the size of the basin of attraction of an equilibrium. (The basin of attraction for a given equilibrium is the set of states that lead to the equilibrium.) A larger basin of attraction should imply that the equilibrium will be attained more often.

More interesting from an empirical perspective, different methods of learning or behavior can lead to a different equilibrium. For example, consider the following simple situation. An agent wants to meet a friend for lunch and knows that they are going to meet at one of two restaurants, A or B, in 10 minutes, but the agent's phone is broken, and he or she cannot contact the friend to coordinate on which restaurant. So the agent chooses one of the restaurants. Suppose that the agent fails to coordinate with the friend, who chose the other restaurant, so the agent eats alone. Now suppose that the next day the same situation occurs. What should the agent do? What strategy does he or she follow? One option is to follow a pure best response to what happened in the previous period. This strategy would lead the agent to choose the opposite restaurant from the last time. If the friend follows the same strategy, they will fail to coordinate again. On the other hand, suppose that the agent plays a best response to the entire history of the friend's choices. The agent chooses according to the fraction of times the friend chooses each restaurant. Thus, as the agent keeps choosing the wrong restaurant time after time, his or her frequency of visiting each one approaches one half. Thus, choosing Restaurant A 50% of the time and Restaurant B 50% of the time, the agent will be able to meet the friend for lunch in 50% of the cases. (One quarter of the time, both choose A; one quarter of the time, both choose B; one quarter of the time, the agent chooses A and the friend chooses B; and one quarter of the time, the agent chooses B and the friend chooses A.) But one could do even better in this case by weighting the more recent choices more strongly.

Then as chance meetings occur, the probability of another chance meeting increases. But recall that one can't go too far. If one plays a pure best response to only the last period, one will return to the possibility of never coordinating. As a side note here, suppose that one prefers Restaurant A and the friend prefers restaurant B. One could also analyze this game from the perspective of an altruist and an egoist strategy. An egoist chooses the restaurant that he or she prefers all the time, and an altruist chooses the restaurant that his or her partner prefers all the time. If two altruists or two egoists (again with conflicting restaurant preferences) play the game, they never coordinate. But if one of each plays the game, they always coordinate. Thus, there can be situations where diversity of preferences, tastes, or types can lead to better outcomes than if diversity is lacking.

Equilibrium selection may also be a function of path dependence (Page, 2006). For instance, in an example given previously, once Operating System X is chosen over Y by a majority of the population, it may be very hard to break out of this equilibrium. Further, even if X and Y start out equal in terms of quality, it may be the case that one comes to dominate the market. But if the minority technology makes improvements and becomes the superior choice, it may be hard to get the population to switch to the better equilibrium. Simply put, history can matter. Further, institutions matter too. As a simple example, there may be a role for the government to help push the public toward coordinating on a superior equilibrium if the population is stuck at an inferior equilibrium. Public policy can be used as a lever to push systems toward or away from one equilibrium or another, depending on the best interest of society.

## Understanding Complex Systems: The Tools

There are a variety of tools used to understand complex systems. Traditional analytical modeling and statistical and econometric techniques are sometimes helpful in understanding complex systems. But their use is often limited by the severe nonlinearity of many of the systems. As mentioned previously, the nonlinearity results from the multiple interdependent relationships and feedback effects common in complex systems models. Because mathematical and statistical methods for dealing with nonlinear systems are limited, computational simulations and particularly agent-based computational experiments are commonly used in the study of complex systems (Miller & Page, 2007).

An agent-based model begins with assumptions about the preferences and behavior of individual agents, firms, and institutions and the interrelationships among them. After specifying the initial conditions (agent endowments of wealth, firm endowments of capital, etc.), a group of artificial agents is set forth and studied. Economists can then vary conditions in the model, changing, say, initial wealth endowments or different learning rules and perform a controlled experiment of the computational system. Such computational models are already common in many natural sciences—physics, for example. Use of these methods in economics is beginning to take hold, and there are a growing number of researchers engaged in agent-based computational economics (ACE). (See Leigh Tesfatsion's Web site for a great overview of the current literature.)

As ACE grows, there are opportunities for complex systems research to merge and collaborate with the equally exciting and emerging field of experimental economics. Experimental economists use populations (usually small) of human subjects to engage in controlled experiments of economic relevance, perhaps risky gambles where individual risk preferences can be uncovered. One limit to human subject experiments is the scale to which the experiments can be run. Typical experiments may have dozens or sometimes a couple hundred participants. An agent-based computational experiment does not suffer from this limit. In fact, it is the goal of one group of economists using agent-based methods to build a model of the entire U.S. economy with one artificial agent for each person and firm residing in the United States. Of course, one strong limit of agent-based models is the fact that the agents are artificial, not real people. Thus, there is fertile ground that can be covered by using agent-based models to scale up human subject experiments and to use human subject experiments to add realism to artificial agent experiments.

## An Example: Schelling's Residential Segregation Model

As an example of a simple complex systems model in economics, consider Thomas Schelling's (1978) residential segregation model. The model is so simple that one may be surprised that it generates segregation at all. To begin, assume that there is a town where all houses are located along one street like the one below:

x x x x x x x x x

Further assume that some of the homes are inhabited by people from the land of p, some are inhabited by people from the land of k, and some of the homes are vacant. The vacant homes are denoted with a v.

k p v v p p k p k

In this example, there are nine homes, seven families (four p families and three k families), and two vacant homes. Now assume that all houses have the same value to all families and that a family can swap its house for a vacant one at any time. Further assume that a family is satisfied with its home as long as one of the neighbors is of the same type as the family; a p is satisfied as long as he or she has one p neighbor, and a k is satisfied as long as he or she has one k neighbor. Thus, in this example, the only two families that are happy are the two p families at the fifth and sixth homes. All the other five families are unhappy about their current living conditions.

Now introduce dynamic movement into the model as follows. Take the citizens in order of location from left to right

and ask each if he or she is satisfied at his or her current location. If the resident is not satisfied, ask if he or she would be satisfied at any of the homes currently vacant. If the resident prefers a vacant home to the current home, move him or her there. If there are two homes the resident would prefer to the current home, move him or her to the home nearest to the current location. If the resident is not satisfied but there is no vacant home in which he or she would be satisfied, the resident waits until the next opportunity to move and checks the vacant homes again. Thus, one would start with the family at the first home. This is a k family with one neighbor, who is a p. The resident is not satisfied with the current home. There are two locations vacant (third and fourth homes), but neither of these locations has a k neighbor. Thus, the resident would be unsatisfied at both of these locations as well and remains at the first home. Now move to the family at the second home. This is a p family with one k neighbor and one vacant neighbor. Since this family does not have a p neighbor, it is not satisfied either. Thus, it will move if it can find a vacant home with a p neighbor. The fourth home fits this criterion. Thus, the family at the second home moves to the fourth. One now has as follows:

k v v p p p k p k

The next three occupied homes (fourth, fifth, and sixth) are all satisfied. Thus, they remain at their current locations. The family at the seventh home is not satisfied since it has two p neighbors. But it could move to the second home and be satisfied. Thus, one now has as follows:

k k v p p p v p k

At the eighth home, there is a p family that is not satisfied. There are two locations this family could move to in order to be satisfied, the third and seventh home, and they move to the closest one, the seventh.

k v k p p p p v k

Finally, there is a k family at the ninth home that is not satisfied. It can move to the third home and become satisfied. This move yields the following:

k k k p p p p v v

If one checks all of the families again, one sees that they are all satisfied. And what one also may notice is that the neighborhood has become completely segregated with three k's and four p's all living next to each other. Note that this high level of segregation occurs even though the families required only that one of their neighbors be similar to them. Even with these modest preferences, the model generated a large amount of segregation.

Note that perfect segregation is not the only outcome in the model. The following configuration would be an equilibrium as well:

p p k k k p p v v

And there are other equilibria similar to this that would not be perfectly segregated. However, what one sees in this simple model is that even modest preferences for wanting to have neighbors similar to oneself can lead to large amounts of homogeneity within neighborhoods and large amounts of segregation.

One can make the model slightly more complex by introducing more dimensions to the neighborhood. Assume that there is a population of individuals that live in Squareville. Squareville is a set of 36 residential locations like the one below:

x x x x x x

x x x x x x

x x x x x x

x x x x x x

x x x x x x

x x x x x x

Again, one can populate the neighborhood with a set of agents of two types and some vacant locations. Again, let the families be satisfied if at least one half of their neighbors are of the same type as them. This time, take each family in a random order and check to see if it is satisfied. If it is, leave the family there. If it is not, look for the nearest location where the family can be happy and move it there. (Break ties with a flip of a coin.)

One can do this in the following way: Take 24 coins and place them on a grid like the one above. Place 12 of the coins on the grid heads up and 12 of the coins on the grid tails up at random locations. Now roll a six-sided die two times. Let the first number be the row and the second number be the column. Thus, if one rolls a two and then a three, look at the coin located at row two and column three. If there is not a coin there, roll again. If there is a coin there, determine if that coin is satisfied or unsatisfied. If it is satisfied, roll again. If it is unsatisfied, find the nearest location where the coin would be satisfied and move it there.

What one will notice is that as one proceeds with this algorithm, patches on the grid begin to develop where there are mostly heads and others where there are mostly tails. And eventually one will reach a point where every coin is satisfied, and in a majority of the cases, there is a very large degree of segregation. Even though each coin would be satisfied if it had only one half of its neighbors like it, many of the coins have only neighbors like them. The grid will have patches of heads only and tails only, with some borders in between.

One can also try this model with different parameters. What happens if families require only one third of their neighbors to be the same as them? What about three quarters? What if there are more vacant spaces? Fewer vacant spaces? What one will find is that the details of the outcomes

will vary (for example, the location of the tails and heads neighborhoods on the grid), but the amount of segregation will still be surprisingly high. For instance, if one sets the tolerance parameter to be one third, significantly more than one third of a family's neighbors will be the same as the family. (In addition to the exercise described, there are several simulation applets available on the Web that can be found in a quick search with such terms as *Schelling segregation model simulation.* One of the simplest to use is the NetLogo Schelling segregation model.)

This simple example displays many of the previously described characteristics of a complex system. The system is dynamic. There are multiple equilibria in the location choices of residents. The coordination on a given equilibrium may have very little to do with the preferences of agents; it may be a product of historical chance. Positive feedback results as neighborhood composition changes. For instance, a decrease in one neighborhood of type p individuals makes it less likely that other type p individuals will remain in the neighborhood. It may not be obvious from understanding the individual incentives and preferences of the agents that large amounts of segregation are likely to result. Simple agent motives lead to complex behavior and outcomes. Neighbors result from a specified interaction structure, in this example a line or a grid, but more complicated structures can incorporate actual neighborhood structures. Finally, even though the model is simple, diversity exists both in the types of agents and in the neighbor of a specified location.

## Conclusion

This chapter has outlined the emerging field of complex systems in economics. All of the hallmark aspects of complex systems are present in almost all economic contexts. Complex systems are generally composed of boundedly rational, diverse agents and institutions who interact within a specified structure in a dynamic environment. Further, these agents and institutions often learn and update their behaviors and strategies and lack a centralized authority that oversees control of the system. These characteristics make the study of economics from a complex systems perspective natural.

Once one embraces a complex systems thinking within economics, many new avenues and tools for research open for one's consideration. There is a rich history in fields such as physics that customarily deal with the nonlinear nature of complex systems models. As such, tools from disciplines such as nonlinear mathematics, computational simulations, and agent-based computational experiments are becoming more and more common in economics. These tools help economists to work through the complicated feedbacks and existence of multiple equilibria that are common in many complex systems.

In addition, it should be recognized that the complex systems approach to economics considers many tools from a variety of approaches that are already common in economics. For instance, learning models are not unusual in contemporary economics, nor are interactions across social networks or diversity. But as with the idea of complexity, the sum of combining many of these constituent parts often leads to a more rich environment and understanding than the analysis of these parts individually.

## References and Further Readings


Anderson, P. W., Arrow, K. J., & Pines, D. (1988). *The economy as an evolving complex system.* Reading, MA: Addison-Wesley.

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *American Economic Review (Papers and Proceedings), 84,* 406–411.

Arthur, W. B., Durlauf, S. N., & Lane, D. A. (1997). *The economy as an evolving complex system 2.* Reading, MA: Addison-Wesley.

Axelrod, R. (1997). *The complexity of cooperation: Agent based models of competition and collaboration.* Princeton, NJ: Princeton University Press.

Blume, L. E., & Durlauf, S. N. (2006). *The economy as an evolving complex system 3.* Oxford, UK: Oxford University Press.

Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies from the bottom up.* Washington, DC: Brookings Institution.

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games.* Cambridge: MIT Press.

Granovetter, M. (1995). *Getting a job: A study of contacts and careers.* Chicago: University of Chicago Press.

Holland, J. (1995). *Hidden order: How adaptation builds complexity.* Reading, MA: Helix Books.

Holland, J. (1998). *Emergence: From chaos to order.* Reading, MA: Helix Books.

Holt, C. A. (2007). *Markets, games and strategic behavior.* Boston: Pearson.

Jackson, M. O. (2008). *Social and economics networks.* Princeton, NJ: Princeton University Press.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71–87.

Miller, J. H., & Page, S. E. (2007). *Complex adaptive systems: An introduction to computational models of social life.* Princeton, NJ: Princeton University Press.

Page, S. E. (2006): Essay: Path dependence. *Quarterly Journal of Political Science, 1,* 87–115.

Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies.* Princeton, NJ: Princeton University Press.

Rosser, J. B., Jr. (Ed.). (2004). *Complexity in economics.* Northampton, MA: Edward Elgar.

Rubinstein, A. (1998). *Modeling bounded rationality.* Cambridge: MIT Press.

Schelling, T. C. (1978). *Micromotives and macrobehavior.* New York: W. W. Norton.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge: MIT Press.

Tesfatsion, L. (2008). *Agent-based computational economics.* Available at http://www.econ.iastate.edu/tesfatsi/ace.htm

Waldrop, M. M. (1992). *Complexity: The emerging science at the edge of order and chaos.* New York: Simon & Schuster.

# 87

# Ethics and Economics

Victor V. Claar
*Henderson State University*

Can markets and market interactions be viewed as ethical? This is a crucial consideration inasmuch as it calls into question the entire discipline of economics and every human action within each market. Markets might very well afford individuals opportunities to choose among a variety of options and to choose rightly. If so, then markets are neither moral nor amoral, yet they create occasions for each individual to make ethical decisions.

But markets are more than this from the viewpoint of the broader society. Although it is true that an individual market participant may be thinking of only his or her own family in deciding whether to purchase a quart of milk, individual participation in market interactions nevertheless leads to outcomes that—over time—lead to the mutual benefit of all, even to those not directly involved in the exchange component of given market trade. This was the point made when Adam Smith conducted his inquiry into what leads to the growing wealth of nations. Though it is no one person's job to make sure that societal wealth and opportunities evolve, improve, and grow over time, market systems accomplish it anyway. And in the view of Smith and his successors, this great feat of the enrichment of humanity can hardly be thought unethical—though individuals may nevertheless be tempted by and succumb to occasional fits of pure selfishness. Moreover, in Smith's view, indeed it is individuals' own self-interest that leads them to behave in ways that are for the betterment of others, inasmuch as they value the esteem of others in their reference groups. So self-interest can reinforce rightful actions, letting individuals measure up in their own eyes as well as in the eyes of others.

Yet critics of markets and market systems accuse markets themselves of rewarding only selfish behaviors. In this brutish view, the most selfish person gets farthest ahead, at the expense of others. And according to this story, a sort of market Darwinism happens wherein those who quickly appreciate that the system rewards the selfish and opportunistic will adapt, behave still more selfishly, and eventually dominate the others as the natural selection process unfolds. This is also why some charge that the study of economics leads to more selfishness: Better understandings and appreciations of markets systems—so the argument goes—lead to more selfish thoughts and actions.

Even in the face of these doubts about markets and the science of their study—economics—most economists are hopeful that markets enrich the lives and work of everyone, including governments, private citizens, nongovernmental organizations, places of worship, and private organizations of all kinds. Indeed, the most rapid path out of poverty and early death is one that passes through a system of law that treats each human being as an equal and one that leads to a fully flourishing system of markets.

To stay on task, this chapter is organized in the following manner. The next section summarizes linkages between economics and ethics prior to Adam Smith, a period influenced by ancient philosophers such as Aristotle, as well as Scholastics trained in the tradition of St. Thomas Aquinas.

Next, the chapter turns to the modern era in economics, using the writing of Adam Smith as the threshold. Building on the tradition of Aristotle and the Scholastics, Smith explicitly linked the interactions of markets to the ethical motivations of human beings, making the case that morals and markets need each other for the good of all members of society.

Following a careful accounting of this period, which persists into the present day, the chapter summarizes critiques of modern economic thinking where ethics is concerned.

Critics of modern economic thinking, on ethical grounds, come from both inside and outside the profession. Outsiders are often theologians and moral philosophers who attempt to knock down an unfair characterization of modern economics, misunderstanding what economics says (and does not say) about human motives and not firmly grasping Smith's position on self-interest. Further, the theologians often claim the moral high ground, inasmuch as they have been to seminary while most economists have not.

Finally, because one of the accusations leveled at modern economics is that studying it leads to more self-interested behavior on the part of its students, this chapter examines this charge. There is limited empirical evidence on this question, though, and the findings are mixed.

## Early Economists and Ethicists

Since the very beginning, economic thinking has been closely linked to ethical thinking. This linkage derives from the intellectual backgrounds and interests of the first economists. Beginning with early philosophers such as Aristotle, the first economic thinkers were also moral philosophers. Centuries later, Scholastics—educated in the tradition of Thomas Aquinas—brought their expertise as moral theologians to the choices people face and the decisions they eventually make. Throughout most of this rich, historical tradition—entwining morals and market considerations—economic decision making and behavior simply were not thought to be in conflict with ethical conduct. On the contrary, for both Aristotle and the Scholastics, human actions and choices within markets make living a good life possible.

### Plato and Aristotle

Little regarding economics and ethics exists prior to Plato (ca. 427 B.C.E.–327 B.C.E.). In his defining and best-known work, *The Republic,* Plato outlines both his understanding of economics as well as his personal insights into the ethical limitations of trade. *The Republic* analyzes both political and economic life, and in it, Plato outlines his views of economy. According to Robert Ekelund and Robert Hébert (1990), though, Plato stopped short of developing any sort of economic theory of exchange; he was more concerned with the resulting distribution of wealth, rather than any specific theory behind trading goods or services. In Plato's view, trading was likely to be a zero-sum game, rather than a means through which the mutual benefit of all might result.

Because Plato saw all profit-seeking activities as potentially corrosive to society, Plato championed a significant role for the state as a regulator and administrator in chief. The state would need to intervene in order to maintain civility, as well as to oversee the resulting distribution.

Though Aristotle was foremost among Plato's students, Aristotle's conception of humanity and its place in the world differed considerably from that of his teacher. For Aristotle, humanity's highest aspiration is to live a good life: a life that is moral and one that thus requires the exercise of virtue.

In his writings, Aristotle outlines a collection of social sciences that he calls *practical sciences,* ones that lead to living a good life. Included among Aristotle's practical sciences are both ethics and economics.

Ricardo Crespo (2008) provides a particularly clear linking of economics, *oikonomike,* and ethics in Aristotle's writings. Where economics is concerned, it may simply be thought of as the study of the ways in which individuals use the things around them in their pursuit of a good life. Inasmuch as individuals take actions in pursuit of a good life, there can be nothing necessarily immoral about such behaviors. In fact, one might regard each action taken in pursuit of a good life as a moral action, inasmuch as pursuing the good life is indeed moral.

To not oversimplify, it is worth considering Crespo's (2006) earlier essay, in which he spells out four different possible meanings for *oikonomike* as used by Aristotle. According to Crespo, *the economic* is actually an analogical term, meaning that it has multiple interpretations, though one of the meanings is central and forms a core around which other related meanings for the same term may be identified. Crespo sees the central meaning of the economic as the human actions taken with the eventual goal of a leading a virtuous—that is, so-called good—life.

Surrounding this central definition of *oikonomike* as human action lie three other orbital definitions, according to Crespo. First, the economic can refer to one's capacity, ability, or talent for the pursuit of good things. Second, the economic can also refer to the habits individuals develop in their pursuit of virtue. That is, through their repeated practice of individual moral actions, each develops a discipline to practice virtue. And third, the economic may also be the practical science of economics itself: the study of human actions in the pursuit of good lives.

As a consequence, then, Aristotle envisioned a much smaller role for the state: one that is limited in scope, predictable, and reliable in relation to the choices made by individuals. A heavy-handed state, such as that suggested by Plato, would steal from each individual the freedom and opportunity to exercise wise judgment in the pursuit of lives that are good. Thus, Aristotle championed private property among all classes of humanity, operating within a market-based economy in which each individual engages in trades that prove beneficial to both parties in a voluntary exchange. In this way, society will become more efficient in its use of resources, enjoy peace and civility, and develop good moral habits.

This is not to say that Aristotle favored no government at all. On the contrary, civil institutions potentially could play two very important roles. First, institutional arrangements could reassure citizens that they were truly equal under the law, regardless of class. An important role of the

state, then, is to establish the rights to private property, as well as to ensure that all property transfers are exclusively voluntary. In this way, society's economic habits are free to emerge, and one will thus be able to study the economic as a practical social science. Second, a predictable and reliable institutional arrangement enhances efficiency. One need only consider any of the modern nations governed by ruthless and arbitrary dictatorial regimes to appreciate that economies do not grow and flourish when the people do not have reasonable assurances about what really belongs to them. Working hard to create even a little wealth might be a terrible bet if the ruling class can take whatever it likes from the rest.

When one puts together Aristotle's thinking on economics and ethics, it is easy to conclude that Aristotle's vision for economic life is one in which actors have sufficient freedom to make their own moral decisions in hopes of living virtue-filled lives. Freedom itself gives each individual constant opportunities to choose well and to develop well-formed habits of doing so.

## The Scholastics and Natural Law

Europe during the Middle Ages bore a much stronger resemblance to the Platonic view of a hierarchical order and state than to the world Aristotle envisioned in which all were equal under the law and the primary role of the state was to establish and maintain the rights to private property. Indeed, Ekelund and Hébert (1990) describe the period as one in which most people belonged to (a) the peasant class, (b) the military, or (c) the clergy. Because it was the clerics who spent time in solitude and contemplation, thinking deep thoughts about the moral order of things, it was these clergy—working throughout a period from roughly 1200 to 1600—who advanced Aristotle's thought linking ethics to economics. As men of faith who were also well-educated men of letters, the Scholastics quite naturally brought their theological and moral thinking to all parts of life.

Ekelund and Hébert (1990) identify five clerics who exemplify the Scholastic tradition that linked Aristotle's economic and ethical thinking to moral thinking in Catholic Europe: Albertus Magnus (ca. 1206–1280), Thomas Aquinas (ca. 1225–1274), Henry of Friemar (ca. 1245–1340), Jean Buridan (ca. 1295–1358), and Gerald Odonis (ca. 1290–1349). Llewellyn Rockwell (1995) extends this list to include late Scholastics from the School of Salamanca. In sixteenth-century Spain, the University of Salamanca served as the center of Scholastic thought. Key economic thinkers from the School of Salamanca include Francisco de Vitoria (1485–1546), Martin de Azpilcueta Navarrus (1493–1586), Diego de Covarrubias y Leiva (1512–1577), and Luis de Molina (1535–1601).

Playing a key role in the linkage of the thinking of the Scholastics to Aristotle is a concept known as *natural law.* Though Aristotle is sometimes viewed as the father of natural law, it is not a term he explicitly used himself. Nevertheless, it is clear that his thinking informed the concept of natural law for the Stoics and later the Scholastics.

The natural law tradition holds that there are certain moral rules or obligations, established in the fundamental nature of things, that apply to all individuals at all times and in all places. These universal truths govern all men and women and cannot be usurped by the laws of humankind. They are inviolable and establish the basic ground rules with which human beings are to conduct themselves in their dealings with each other. In this tradition, all men and women are equal under the natural law. One could think of natural law as an undeniable, unavoidable force like gravity: Regardless of wealth or lineage, if one jumps from a tall building, one is just as likely as anyone else to get seriously injured.

To further illustrate the notion of natural law, it is helpful to refer to the famous second sentence of the Declaration of Independence of the United States. Its authors were surely writing in the natural law tradition when they wrote, "We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness." For the framers of the Declaration, it required no argument that "all men are created equal" or that each individual is born with some fundamental human rights that cannot be sold or taken away. These claims were thought to be entirely obvious to all: They were "self-evident."

Natural law ideas are clearly present in the writing of Aristotle, though not explicitly stated. Instead, Aristotle contrasts particular laws that might be written by humankind with common laws that have bearing everywhere. And certainly the equality of humans and their right to private property in the pursuit of a moral life may be viewed as part of the common—rather than particular—law.

Because the Scholastics served as the moral umpires for the church during the period, their main interest in economics derived largely from its potential to distribute justice. That is, the Scholastics were more interested in the potential for justice following from economic actions than they were in any specific details about exchange mechanisms. So they employed their reason in the search for universal, transcendent ideas that govern everyone. And certainly economic laws govern individuals' interactions with each other and also apply equally to all people in all places. So the Scholastics' study of economics was driven by their desire to discern universal truths—natural law—that give insight into the nature of the universe in which individuals live and interact. Of course, although Aristotle did not ascribe the origin of natural law to a higher being, it was easy and obvious for the Scholastics to read Aristotle and see God as the ultimate source of the natural law under which everyone lives, and economic forces as part of God's natural law. So the Scholastics studied economics in order to better understand God's created order.

Given that the Scholastics followed in the natural law tradition begun by Aristotle, they also viewed all humans as equal, both in the sight of God and in the sight of each other. And viewing economic forces as ones that are blind to the specific plight of a given person, the Scholastics followed Aristotle's lead and viewed economic decisions and actions as opportunities to act morally. Thus, like Aristotle, the Scholastics viewed the proper role of civic institutions as a limited one designed to defend, predictably and reliably, the fundamental rights of all.

Rockwell (1995) gives a few specific examples of the economic thought of the late Scholastics, illustrating how the work of the School of Salamanca clearly follows from Aristotle and also anticipates well the modern study of economics that begins with Adam Smith. Each is especially intriguing in light of modern antipathy from theologians toward economics and economists, which is discussed later in the chapter.

According to Rockwell (1995), Luis de Molina was among the first of the Scholastics in the Jesuit order to think very carefully about theoretical topics in economics. In his defense of private property, he simply pointed to the commandment that says, "Thou shalt not steal." But Molina went farther and began to explore modern economic arguments for private property, including what is now called the *tragedy of the commons:* that when everyone owns a resource, then no one has the correct incentive to care for it. Good stewardship of a resource comes only when the steward is also the owner; only then are incentives correctly aligned. Molina also saw government—the king—as capable of great moral sin against the good of the people when government's powers are either broad or arbitrarily applied.

Covarrubias took private property ownership to the extreme, making the case that the owner of a piece of land is the only person who has the moral right to decide what to do with its fruits. In fact, Covarrubias claimed that government could not intervene even if a landowner were growing some kind of medicine but refused to sell. If the landowner were forcibly made to do something against his or her will with property he or she rightly owned, even if it might be beneficial to others, such an action would constitute violation of the natural law.

## Orthodox Views of Ethical Behavior Within Markets

### Adam Smith, *Wealth of Nations,* and *Theory of Moral Sentiments*

Modern thought on ethics and economics begins with the work of Scottish professor Adam Smith (1723–1790). All students of economics should spend at least some time reading Smith in order to see for themselves what Smith says—and what he does not say—regarding self-interest, selfishness, and the market. In particular, students should be sure to look closely at Smith's best-known work, *An Inquiry Into the Nature and Causes of the Wealth of Nations* (1776/1981), as well as his earlier *Theory of Moral Sentiments* (1759/1982). As this chapter shows, Smith did not view morals and markets as conflicting forces. On the contrary, Smith viewed market behavior and moral behavior as cooperative activities, following from similar views regarding what motivates each individual to action or inaction. Although many practicing modern mainstream economists have forgotten much of Smith's thinking on ethical behavior, remembering instead only his treatment of self-interest in *The Wealth of Nations,* modern economics nevertheless gets exactly right the Smithian conclusion regarding markets and their tremendous potential for improving the good of all.

Smith represented the synthesis of the natural law tradition begun by Aristotle, then articulated by the Stoics, the Scholastics, and later the French Physiocrats, following the tradition of their countryman and forebear Pierre le Pesant de Boisguilbert. The Physiocrats, led by François Quesnay, had believed that the natural law served as a reflection of the creator. In their view, then, natural law needed to take precedence over laws created by humankind. Further still, the laws of humans were certainly not as good as the natural laws of the creator. After all, the creator is the Supreme Being. In this light, the Physiocrats thought that the laws enacted by humans (called *positive law*)—flawed as they are—should be kept to a minimum. Society would be much healthier and in greater harmony with the mind of the creator if it learned to lean on natural law more and on positive law less.

In this light, the stage had been set for the laissez-faire (natural liberty) views articulated by Smith in his writings. And to grasp clearly how Smith integrated ethics, economics, and natural law to form a seamless natural theology—which owes much indeed to the natural law tradition—one must consider together *The Wealth of Nations* and *The Theory of Moral Sentiments.* To focus on only *The Wealth of Nations,* which is certainly what most economists do (if they read it at all), is to miss the primary articulation of Smith's natural theology. Indeed, *The Theory of Moral Sentiments* is a theory of how and why individuals each act in a largely moral way in their affairs. *The Wealth of Nations* is merely an extension of Smith's theory to individuals' interactions in markets and the power of each of those actions to have a cumulatively beneficial impact on society as a whole. Or to state it another way, *The Theory of Moral Sentiments* is Smith's statement of his ethics; *The Wealth of Nations* is the application of Smith's ethics to economics.

But this discussion begins the way most students of economics begin: first with *The Wealth of Nations* (henceforth *WN*), then working back to Smith's earlier *Theory of Moral Sentiments* (henceforth *TMS*). In looking at *WN,* the main goal is to grasp accurately what Smith said—and did not

say—about the role of self-interest (or self-love) in *WN* and whether it is indeed ethical to afford a role to self-interest in people's economic dealings with each other. Then, in looking at *TMS,* this chapter attempts to work out what has been referred to as the *Adam Smith problem:* that the same fellow who articulated the role of self-interest in economic dealings had put forward an entire theory of our moral feelings and actions just 17 years earlier.

Although much of *WN* is widely known (e.g., the pin factory as an illustration of the production efficiency gains possible from specialization of tasks and the division of labor), the concern here is with only the passages relevant to the relationship between ethics and economics. The most famous idea relevant to the present purpose is the role that self-interest plays in market dealings. And unfortunately, the most common telling of the story of self-interest misses much of Smith's point. To clarify, this chapter first reviews this common telling that mischaracterizes Smith's articulation of the role of self-interest in markets.

The common story proceeds like this. Economic agents—people and firms—are self-interested; they like to be happier (or wealthier or both). Seeking greater happiness, they pursue opportunities that leave them happier than they currently are. And in a decentralized market system with no central planner, potential trading partners will discover each other and execute trades that leave them both better off. These mutually beneficial exchanges improve the lot of both traders; if this were not the case, they would not trade, because trade is voluntary. Also, no one—neither the traders nor any bystander—is harmed as a result (in most circumstances). Much as in Aristotle's view, each person is at liberty to pursue a personal course of action leading to a good life as he or she sees it, and mutually beneficial trades are small steps making the good life possible. Even greater still, when many such actions are repeatedly taken over time, across an entire society, the end result is a better life for all. Indeed, markets and their workings help build the wealth of nations, even though no central planner or planning committee is in charge.

What this story leaves out is what Smith actually wrote in *WN.* True, Smith did make the final point in the preceding paragraph: Decentralized, mutually beneficial exchange ultimately grows the wealth of nations. But a common misunderstanding (or misrepresentation) of Smith is the role that self-interest plays in a market transaction. To consider this more carefully, look at the key passage in *WN,* found in Chapter 2 of Book 1. Smith writes,

> In civilized society [man] stands at all times in need of the co–operation and assistance of great multitudes, while his whole life is scarce sufficient to gain the friendship of a few persons. In almost every other race of animals each individual, when it is grown up to maturity, is intirely independent, and in its natural state has occasion for the assistance of no other living creature. But man has almost constant occasion for the help of his brethren, and it is in vain for him to expect it from their benevolence only. He will be more likely to prevail if he can interest their self-love in his favour, and shew them that it is for their own advantage to do for him what he requires of them. . . . It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages. (Smith, 1776/1981, pp. 26–27)

Reading this passage makes clear what many—including many economists—misunderstand about Smith's view of the role of self-interest. Suppose, for example, that one feels a strong need to acquire some pencils in the progress toward a good life. Without market exchange, how could one get them? Well, one could try to make some pencils, but as Leonard Read (1958/1999) makes thoroughly clear, one person working alone might not be able to complete even one pencil in a lifetime, if one must rely on no one but oneself for every single input and production step required.

Assuming one needs pencils relatively soon, or at least before death so that they can be of some help in a good life, what other possibilities are there? A second avenue might be to become great friends with someone who currently possesses some pencils and hope that the friend's affection is so great that he or she will simply give away some pencils. But again, working at sufficiently ingratiating oneself with another person so that he or she will give one gifts requires time, effort, and energy. Moreover, attempting to acquire all of the goods and services one needs throughout one's life in this way—appealing to the good nature of others in the hope of receiving gifts—is a dubious undertaking indeed.

A third possible strategy for obtaining pencils would be simply to beg for them—and to beg for everything else one needs for a good life. One could go around town on bended knee, with a tin cup in hand, and beg, attempting to convince anyone one meets that one's need for pencils is more pressing or more important than the pencil owner's needs are. But as Robert Black (2006) points out, begging for what one needs would clearly violate the idea of natural law. Begging is dehumanizing and reduces the value and sense of self-worth of the beggar. One who has even a single experience of begging in one's life probably remembers how miserable it felt to be forced by life's circumstances into that situation. In the natural law tradition, all human beings are of equal value, regardless of class, and are entitled to maintain their dignity as human beings.

A fourth option is force. One could simply steal pencils. But that option falls short because it involves taking another person's property, which is clearly at odds with the natural law tradition.

Having exhausted (a) making pencils, (b) ingratiating one's self with others who have pencils, (c) begging for pencils, and (d) stealing pencils, what remains? For Smith, only one possibility remained that maintained the dignity of all, and it also smoothly and quickly moved goods and services from less- to more-valued uses. If a person wants

pencils, it might be simplest for him or her to appeal to the self-interest (self-love) of those who have pencils or make pencils and convince them how much better off they would be if they traded away some of their pencils. That is, a highly efficient way to get others to do things one wants, which also lets everyone keep their dignity, is to appeal to others' self-interest. If individuals do this well, then others will happily trade with them, thereby making the individuals' lives happier and easier also.

Simply put, it is not mere selfishness that makes one better off in Smith's view; it is one's appeal to the self-interest of others. In fact, everyone's lives depend on it.

Having clarified what Smith said and did not say in *WN* regarding self-interest, this chapter turns to the reconciliation of *WN* with Smith's earlier *TMS.* As discussed previously, scholars have struggled for many years with the Adam Smith problem: reconciling Smith's moral theory in *TMS* with the role of self-interest in *WN.*

James Otteson (2002) provides one of the clearest answers to the Smith problem. For Otteson, both *TMS* and *WN* share a common theme: Natural law and self-love work together as a powerful force to explain and predict human behaviors and interactions, whether those interactions are in the marketplace or in social settings. Either way, even with no one person in charge of either markets or morality, there is a natural ordering that results when men and women everywhere have the freedom to live a good life. In markets, people engage in trades that prove mutually beneficial to both parties. And in relationships with each other, people learn to act in ways that are humane and that preserve the dignity of everyone.

By what mechanism does one discover what is appropriate in social contexts? Smith argues in *TMS* that everyone craves sympathy of feeling from others. People want others to share in their joys, sorrows, pleasures, and pain. People want to be viewed with dignity at all times and to be taken seriously as fellow human beings. So Smith makes the case that just as individuals and firms can learn through market interactions which goods and services are valuable in the marketplace, individuals also learn, through social interactions, which behaviors will be rewarded with sympathy of feeling in the social realm.

Consider two specific examples from *TMS,* in which someone might misunderstand how to express appropriately his or her feelings to others in a social setting: someone who wails over some small hurt and another who laughs inappropriately long at a joke. In both cases, Smith (1759/1982) suggests that others will like the person less because the behaviors are out of proportion to the circumstance:

> If we hear a person loudly lamenting his misfortunes, which, however, upon bringing the case home to ourselves, we feel, can produce no such violent effect upon us, we are shocked at his grief; and, because we cannot enter into it, call it pusillanimity and weakness. It gives us the spleen, on the other hand, to see another too happy or too much elevated. . . . We are even put out of humour if our companion laughs louder or longer at a joke than we think it deserves; that is, than we feel that we ourselves could laugh at it. (p. 16)

What hope is there, then, for each of us—even social clods—to learn how to behave appropriately in society? How can one learn to behave in ways that others will find appropriate yet remain true to one's own inalienable rights? Smith's answer is that throughout people's entire lives, even when they are young, they all are students at the school of self-command. And people are in class at the school of self-command anytime they find themselves interacting with other human beings. So for example, if people are in the bad habit of laughing too long at a joke, presumably they will catch on from the feedback they receive (eye rolling, fewer social invitations, gentle admonitions from friends) about their inappropriate behaviors and modify them, throughout their lives, accordingly.

> A very young child has no self-command; but . . . [w]hen it is old enough to go to school, or to mix with its equals, it . . . naturally wishes to gain their favour, and to avoid their hatred or contempt. Regard even to its own safety teaches it to do so; and it soon finds that it can do so in no other way than by moderating, not only its anger, but all its other passions, to the degree which its play-fellows and companions are likely to be pleased with. It thus enters into the great school of self-command, it studies to be more and more master of itself, and begins to exercise over its own feelings a discipline which the practice of the longest life is very seldom sufficient to bring to complete perfection. (Smith, 1759/1982, p. 145)

And who is to help give instruction and guidance to each individual? Smith (1759/1982) envisioned that each person has a conscience, an "impartial spectator" who recalls all of the lessons one has learned during study at the school of self-command and holds one accountable for all of those lessons:

> The man of real constancy and firmness . . . has never dared to forget for one moment the judgment which the impartial spectator would pass upon his sentiments and conduct. He has never dared to suffer the man within the breast to be absent one moment from his attention. With the eyes of this great inmate he has always been accustomed to regard whatever relates to himself. This habit has become perfectly familiar to him. He has been in the constant practice, and, indeed, under the constant necessity, of modelling, or of endeavouring to model, not only his outward conduct and behaviour, but, as much as he can, even his inward sentiments and feelings, according to those of this awful and respectable judge. He does not merely affect the sentiments of the impartial spectator. He really adopts them. He almost identifies himself with, he almost becomes himself that impartial spectator, and scarce even feels but as that great arbiter of his conduct directs him to feel. (pp. 146–147)

Thus, just as market order is spontaneously driven by people's appeals to others' self-love, a mutually beneficial

social system can result spontaneously as long as individuals crave sympathy of feeling with others and desire to hold personal views that are in accord with those of others. Barring the occasional clod or sociopath who simply cannot learn from the school of self-command, humans learn from each other how to behave appropriately and, more to the point, they learn how to behave morally in regard to others, while nevertheless doing so out of their own self-love: their desire to be liked and esteemed by others. Further, their personal impartial spectator is ever present, reminding them of all of the lessons they have studied over the years at the school of self-command and exhorting them to choose rightly in all things.

Putting *WN* together with *TMS* leads to a fascinating solution to the Adam Smith problem: Smith's overarching view is that human interactions, though motivated by self-love, can nevertheless lead to thoroughly pleasing outcomes, whether the outcomes are economic or social. As long as the dignity of humankind is preserved in the freedom to make choices, individuals will all ultimately be motivated to actions that not only benefit themselves but also benefit the other individuals they know and encounter; with time, society's overall economic wealth will grow, as will its capacity for great moral good.

### Ethics and Economics Since Adam Smith

Even though economics has deepened in terms of economic understanding and broadened in terms of the range of topics it explores, economics has never tossed out Smith's famous invisible hand: By constantly directing resources from less- to more-valuable uses, market forces lead to the eventual betterment of all, even though all individuals are primarily focused on living good lives for themselves, their families, and their friends.

If there is tension in modern economics along ethical lines, it comes mainly from the distinction between efficiency and fairness in the allocation of resources. Efficiency in the allocation of resources refers to *Pareto efficiency,* named for Italian economist and sociologist Vilfredo Pareto (1848–1923). According to the Paretian standard of efficiency, economists describe an allocation of resources as efficient when it is impossible to make one person better-off without harming someone else in the process. Economists are always on the lookout for Pareto-improving possibilities: situations in which one or more persons may be made better-off (in whatever sense that means to them) at no expense to others.

The Paretian standard of efficiency is not in conflict with the market system described by Smith, because an obvious way to make two people better-off without harming others is to let the two people execute any voluntary trades they like. Once all possible potentially Pareto-improving trades in a society have been exhausted, then there will be no way to improve the lot of one without causing another to surrender something involuntarily. Because the Paretian standard is difficult to argue against on moral grounds and because it is consonant with Smith's self-love and natural law ideas, Pareto efficiency is the primary way in which allocations of resources are assessed in welfare economics.

Yet some argue that in its simplicity, Pareto efficiency leaves something out: fairness. That is, they contend that market outcomes, even if Pareto efficient, may not be fair if the outcomes are too unequal. So economists working in welfare economics consider alternative measures of societal welfare besides the Paretian standard. The two most common are utilitarianism (following the work of Jeremy Bentham and other utilitarian thinkers) and Rawlsianism (named for Harvard philosophy professor John Rawls). Because these two alternative measures of society's well-being lie beyond the scope of this chapter, they are not discussed in detail here. Nevertheless, it is important to note that though drawn from different ideological starting points, both utilitarianism and Rawlsianism yield the same clear policy recommendation: Regardless of the unfettered market outcome, all resources need to be redistributed until each member of society is equally well-off. To do otherwise would not be fair. Of course, because such redistributions (a) reduce incentives for high-income earners to create more jobs that lead to greater societal wealth (i.e., a slower growing pie), (b) may prove very costly to administer and control, and (c) encroach on the rights of individuals to maintain private property and surrender it voluntarily, most economists prefer efficiency to fairness as an assessment tool. If nothing else, the Paretian standard is clearly defined. In contrast, fairness is always murky; what appears fair to one usually depends on where one happens to be sitting.

## Modern Suspicions and Critiques of Ethical Views in Economics

Owing largely to the shifting focus from Smith's companion views of morals and markets to one centered on mere human self-interest—frequently misinterpreted as selfishness—modern critiques of economics have come from both within the profession and without. Most critics who are economists try to gently point out that the profession needs to expand its focus to remember what Smith really said. They care deeply and passionately about the great issues of our day, especially global poverty, the environment, and basic human rights. Amartya Sen and William Easterly represent the very best of this movement in modern economics.

But the harshest attacks on economics come from outside the profession. These attacks are mounted on ethical grounds, charging that the models, theories, and policy prescriptions of the social science are driven by a view that selfish, greedy behavior is both perfectly acceptable and to be expected. Stated another way, these critics are somewhat guilty of setting up a straw-man argument: critiquing a caricature of modern economics, rather than economics

as economists actually think about it and talk about it among themselves. And these critics attempt to position themselves as coming from morally high ground—perhaps from out of theology or from the environmental movement.

The remainder of this section considers briefly the interactions of theologians with economics since Smith. And as this section shows, the antipathy of theologians toward economics and economists is a relatively recent phenomenon. Until at least the mid-nineteenth century, clergymen—at least in the United States—embraced economics and its study as a powerful tool to uplift humanity. The late Paul Heyne (2008) gives an excellent overview of this period in Part 4 of *Are Economists Basically Immoral?*, a posthumously published volume of his works.

According to Heyne (2008), in the years before 1800, Christian thinkers had not fully integrated economics and their faith. True, Smith had followed in the natural law tradition of the Christian Scholastics in his thinking on economics and the moral order. Yet Smith was hardly more than a deist: someone who believes that a supreme being created the universe and its natural laws, set the universe in its course, yet left humankind to do what it will.

But the publication of the Reverend Francis Wayland's (1837) *Elements of Political Economy* united the laissez-faire economics of Smith with a distinctly Christian articulation of the natural-law tradition. Wayland's book, which became the most widely used economics textbook in the United States on its publication, echoed *WN*. Wayland claims that the rules for wealth accumulation were part of God's providence (a gift) and that individuals who honestly worked toward improving their own lots would thereby promote the welfare of humanity. Heyne (2008) argues that Wayland and other Christian thinkers were inclined to link the optimism of Smith to their own faith inasmuch as markets gave great promise for dramatic improvement of the human condition. And Heyne considers the Reverend Arthur Latham Perry, professor of history and political economy at Williams College, as the clergy's foremost defender of laissez-faire economics during the period. Indeed, when the American Economic Association was founded in 1885, several Protestant clergy were among its charter members.

Yet as the century drew to a close—and as the association quickly shook off its Christian roots—the clergy quickly grew critical of the promise of markets to do much of the heavy lifting in the transformation of society. Such a quick reversal, from an embrace of economic thinking by the clergy to full-throated criticism of it, is indeed puzzling. By way of explanation, Heyne (2008) posits that it is no coincidence that academic clergy began to grow tired of Smithian economics at the same time that the surrounding academic culture began to grow critical as well. For Heyne, prevailing political and academic cultural winds, far more than any deeply informed change of course, were the main reason that academic clergy moved from fans to foes of the promise of markets.

Lamentably, this division between economics and theologians persists into the present day, and finding theologians who are critical of modern economics is not a difficult task. And more often than not, their criticisms are likely to be founded on a caricature of economics, rather than a richer depiction of it. For example, William Cavanaugh (2008) writes this regarding scarcity:

> The standard assumption of economists that we live in a world of scarce resources is not based simply on an empirical observation of the state of the world, but is based on the assumption that human desire is limitless. In a consumer culture we are conditioned to believe that human desires have no end and are therefore endless. The result is a tragic view of the world, a view in which there is simply never enough to go around, which in turn produces a kind of resignation to the plight of the world's hungry people. (p. xii)

Similarly, Christine Hinze (2004) states, "When they reduce the meaning and purpose of 'economy' to market exchange, or human agents to self-interest-maximizing Homo economicus, mainstream economists fall prey to a category mistake bound to distort both analysis and policy recommendations" (p. 172).

And D. Stephen Long (2000) claims, "As a discipline, economics has increasingly developed an anti-humanistic mode. . . . [E]conomics has become an increasingly abstract—mathematical—science" (p. 9). For two spirited book-length debates between market advocates and market critics, see Doug Bandow and David Schindler (2003) and Donald Hay and Alan Kreider (2001).

Yet many of today's thinking clergy continue to have tremendous faith in markets both as mechanisms that can transform the lives of the global poor living in material poverty and mechanisms that afford individuals a wealth of occasions in which to choose to act morally. Particularly notable is the work of Father Robert Sirico, president of the Acton Institute in Grand Rapids, Michigan, as well as that of Michael Novak (1982), author of *The Spirit of Democratic Capitalism.* Though not theologians, Victor Claar and Robin Klay (2007) defend—from a faith perspective—the power of markets to effect tremendous good.

## Does Studying Economics Change One's Ethics?

A final intersection of ethics and economics concerns whether instruction in economics leads students and eventual economists to think or act in more self-interested ways, as some of the profession's critics might assume. First, there does seem to be some evidence, including a study by Robert Whaples (1995), that taking even an introductory economics course influences students to view market outcomes as more fair. Whaples finds the strongest change in attitudes among women.

Second, empirical evidence concerning whether studying economics leads to changes in one's behavior is mixed. For example, using laboratory-based experiments, John Carter and Michael Irons (1991) find little actual change in behavior between freshman and senior economics students, suggesting that studying economics did not change them. If so, perhaps students who gravitate to the study of economics are already different from other students when they arrive. This finding has been reinforced more recently by Bruno Frey and Stephan Meier (2005) and by Meier and Frey (2004). This more recent work calls into question the earlier finding of Robert Frank, Thomas Gilovich, and Dennis Regan (1993) that economics majors were less likely to cooperate in prisoners' dilemma games than students in other majors. Further, Harvey James and Jeffrey Cohen (2004) find that including an ethics component in an economics program can reduce the tendency to not cooperate in laboratory games.

Nevertheless, there is considerable hope that students and practitioners of economics may perform at least as nobly as others in settings that are not known to be experimental. For example, one of the findings of Carter and Irons (1991) is that economics students stated they would be more likely to keep money that had been lost. Yet in a follow-up experiment using actual money, Anthony Yezer, Robert Goldfarb, and Paul Poppen (1996) discover that students of economics were significantly more likely to return money they had found. In the experiment, $10 was put inside stamped, addressed envelopes, and then the envelopes were dropped into economics classrooms, as well as classes in other disciplines. To return the money, all that was required was to seal an envelope and mail it. Though only 31% of envelopes were returned in business, history, and psychology classes, economics students returned 56% of the envelopes. And David Laband and Richard Beil (1999) find that practicing economists were less likely to cheat on their professional membership dues than professional political scientists and—especially—professional sociologists, suggesting practicing economists behave significantly more ethically than other professional social scientists.

## Conclusion

To this day, Adam Smith's moral theory advanced in *The Theory of Moral Sentiments* and applied to economic interactions in his later *Wealth of Nations* constitutes the foundation and nucleus of all moral and ethical thinking in economics. Barring instances of market failure, market-based exchanges lead to Pareto improvements, because some members of society are made better-off through market exchange, and none are harmed as a result. Further, markets are the most powerful mechanism available for the perpetual redirection of society's resources from less- to more-desired uses, leading to eventual transformation of entire nations as the resulting wealth accumulation leads to a higher standard of living for all. And the personal liberty available in a market system gives each individual ongoing opportunities to make wise moral and ethical choices. Moreover, in a market-based economy, the fastest way for individuals to make life better for themselves and their families is to meet the needs of others. Indeed, perhaps even without human feeling for another, each individual will nevertheless act humanely toward another, because the very means by which people enrich their own lives is by enriching the lives of others. And according to Smith, such relationships are borne out in the social realm as well: People behave decently to others in order to earn others' respect and sympathy of feeling in exchange.

Though most modern theologians are critical of economics and economists, instead hoping for the arrival of a more so-called moral allocative system than the one Smith envisioned, their criticism is a relatively recent phenomenon. Even 400 years ago, Scholastics were working out an economic theory, in the belief that economic laws were among the natural laws given by their creator. They studied economic laws in order to better understand and appreciate their world. Only since the close of the eighteenth century have academic clergy grown harshly critical of the wisdom and ethics of market economics and economic science, perhaps owing to the influence of similar suspicions growing contemporaneously in other parts of academe. Yet the work of market-minded modern theologians such as Michael Novak and Father Robert Sirico of the Acton Institute have begun to renew the faith of the clergy in the power of markets to bring dramatic transformation of society, especially where global poverty is concerned.

Students of economics may perhaps be less cooperative than others, but the empirical evidence is mixed. Further, students selecting economics may already be different from others, suggesting that any perceived differences may be attributable to a selection effect, rather than to any specific treatment effect of studying economics. Further, in nonexperimental settings, recent empirical evidence suggests that both students of economics and professional economists behave more ethically than their counterparts from other disciplines.

## References and Further Readings

Bandow, D., & Schindler, D. L. (Eds.). (2003). *Wealth, poverty, and human destiny.* Wilmington, DE: ISI Books.

Black, R. A. (2006). What did Adam Smith say about self-love? *Journal of Markets & Morality, 9,* 7–34.

Broome, J. (1999). *Ethics out of economics.* Cambridge, UK: Cambridge University Press.

Carter, J. R., & Irons, M. D. (1991). Are economists different, and if so, why? *Journal of Economic Perspectives, 5*(2), 171–177.

Cavanaugh, W. T. (2008). *Being consumed: Economics and Christian desire.* Grand Rapids, MI: William B. Eerdmans.

Claar, V. V., & Klay, R. J. (2007). *Economics in Christian perspective: Theory, policy and life choices.* Downers Grove, IL: IVP Academic.

Crespo, R. F. (2006). The ontology of "the economic": An Aristotelian perspective. *Cambridge Journal of Economics, 30,* 767–781.

Crespo, R. F. (2008). Aristotle's science of economics. In I. R. Harper & S. Gregg (Eds.), *Christian theology and market economics* (pp. 13–24). Cheltenham, UK: Edward Elgar.

Ekelund, R. B., & Hébert, R. F. (1990). *A history of economic theory and method* (3rd ed.). New York: McGraw-Hill.

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). Does studying economics inhibit cooperation? *Journal of Economic Perspectives, 7*(2), 159–171.

Frey, B., & Meier, S. (2005). Selfish and indoctrinated economists? *European Journal of Law and Economics, 19,* 165–171.

Graafland, J. J. (2007). *Economics, ethics, and the market: Introduction and applications.* New York: Routledge.

Harper, I. R., & Gregg, S. (Eds.). (2008). *Christian theology and market economics.* Cheltenham, UK: Edward Elgar.

Hay, D. A., & Kreider, A. (Eds.). (2001). *Christianity and the culture of economics.* Cardiff, UK: University of Wales Press.

Heyne, P. (2008). *"Are economists basically immoral?": And other essays on economics, ethics, and religion* (G. Brennan & A. M. C. Waterman, Eds.). Indianapolis, IN: Liberty Fund.

Hinze, C. F. (2004). What is enough? Catholic social thought, consumption, and material sufficiency. In W. Schweiker & C. Mathewes (Eds.), *Having: Property and possession in religious and social life* (pp. 162–188). Grand Rapids, MI: William B. Eerdmans.

James, H. S., Jr., & Cohen, J. (2004). Does ethics training neutralize the incentives of the prisoner's dilemma? Evidence from a classroom experiment. *Journal of Business Ethics, 50,* 53–61.

*Journal of Markets & Morality: Scholarship for a humane economy.* (n.d.). [Various issues].

Knight, F. H. (1997). *The ethics of competition.* New Brunswick, NJ: Transaction. (Original work published 1935)

Krelle, W. E. (2003). *Economics and ethics 1.* Berlin, Germany: Springer.

Laband, D. N., & Beil, R.O. (1999). Are economists more selfish than other "social" scientists? *Public Choice, 100,* 85–101.

Long, D. S. (2000). *Divine economy: Theology and the market.* London: Routledge.

Meadowcroft, J. (2005). *The ethics of the market.* New York: Palgrave Macmillan.

Meier, S., & Frey, B. (2004). Do business students make good citizens? *International Journal of the Economics of Business, 11,* 141–163.

Novak, M. (1982). *The spirit of democratic capitalism.* New York: Touchstone.

Nozick, R. (1974). *Anarchy, state, and utopia.* New York: Basic Books.

Otteson, J. R. (2002). *Adam Smith's marketplace of life.* Cambridge, UK: Cambridge University Press.

Read, L. E. (1999). I, pencil: My family tree as told to Leonard E. Read. *Library of Economics and Liberty.* Retrieved June 26, 2009, from http://www.econlib.org/library/Essays/rdPncl1.html (Original work published 1958)

Rockwell, L. H. (1995, September). Free market economists: 400 years ago. *Freeman, 45*(9). Retrieved June 5, 2009, from http://www.thefreemanonline.org/featured/free-market-economists-400-years-ago

Roth, T. P. (2002). *The ethics and the economics of minimalist government.* Northampton, MA: Edward Elgar.

Sacks, J. (1999). *Morals and markets.* London: Institute of Economic Affairs.

Sen, A. (1988). *On ethics and economics.* Malden, MA: Blackwell.

Smith, A. (1981). *An inquiry into the nature and causes of the wealth of nations.* Indianapolis, IN: Liberty Fund. (Original work published 1776)

Smith, A. (1982). *The theory of moral sentiments.* Indianapolis, IN: Liberty Fund. (Original work published 1759)

Tamari, M. (1987). *"With all your possessions": Jewish ethics and economic life.* New York: Free Press.

U.S. Catholic Bishops. (1986). *Economic justice for all: Pastoral letter on Catholic social teaching and the U.S. economy.* Retrieved June 5, 2009, from http://www.osjspm.org/economic_justice_for_all.aspx

VanDrunen, D. (2006). *A biblical case for natural law.* Grand Rapids, MI: Acton Institute.

Wayland, F. (1837). *The elements of political economy.* New York: Leavitt, Lord.

Whaples, R. (1995). Changes in attitudes about the fairness of free markets among college economics students. *Journal of Economic Education, 26,* 308–313.

Yeager, L. B. (2001). *Ethics as social science: The moral philosophy of social cooperation.* Cheltenham, UK: Edward Elgar.

Yezer, A. M., Goldfarb, R. S., & Poppen, P. J. (1996). Does studying economics discourage cooperation? Watch what we do, not what we say or how we play. *Journal of Economic Perspectives, 10*(1), 177–186.

# 88

# Feminist Economics

Gillian Hewitson
*University of Sydney*

During the nineteenth and early twentieth centuries, in Western countries such as the United States and Britain, women's economic situation was largely dictated by the legal framework within which economic activity took place. Some examples include the legal requirement, embedded within the common law of marriage, for wives to provide housework, childrearing, and sexual services to their husbands, in return for at least a minimal subsistence. Husbands were legally entitled to the wage earnings of working wives. Wives were, in effect, legal chattels. Women's labor supply decisions were restricted by "protective legislation" with respect to total hours worked, which hours, and in which occupations. Sex discrimination was legal, and women's wages were determined not by their productivity but by socially accepted norms, such as the widespread belief that women worked for "pin money" rather than for economic necessity, which allowed employers to pay them badly. Indeed, working-class women were frequently unable to support themselves, making marriage an attractive economic proposition. Furthermore, women had limited opportunities, if any, to attend university, and they could not own property in their own names. Despite the restrictions of marriage, then, most women married because their access to economic independence was so limited. Women were assumed to be actual or future wives, and the legal and economic environment ensured that this would be so. Even in the mid-twentieth century, women in certain occupations were terminated if they got married; if a wife had a bank account in her own name, her husband had access to it, and a married woman could not borrow money without her husband's consent.

However, by the 1970s, most, if not all, of the legal restrictions on women's economic activity mentioned above had been eliminated in modern Western countries. Most countries have ratified the Convention on the Elimination of Discrimination Against Women. And now there are laws in place that make it illegal to discriminate against women in employment, hours, earnings, and lending. Indeed, the first piece of legislation signed by President Obama was the Lily Ledbetter Act, which widens the scope of women's access to the courts upon discovering wage discrimination on the basis of sex. Wives own their own earnings. Women have become prime ministers and presidential candidates, and they have entered the boardrooms of *Fortune* 500 companies. And, in the United States, women make up nearly half the labor force. This environment, in which women appear to have the same ability as men to determine their own economic fates, sometimes makes young people question the relevance of feminist economics in today's world. As this chapter shows, however, the economic system remains a gendered system, and most women's economic outcomes are related to the fact that they are women. This is true both domestically in the United States and in other Western countries, as well as globally.

Feminist economics is the area of research and practice within which the gendered economic system is analyzed.[1] Feminist economists argue that gender is central to understanding the allocation of economic opportunities, rewards, and punishments, and therefore gender is a key determinant of individual economic outcomes. They use a variety of feminist perspectives to understand and change the social and economic institutions and policies that reinforce the economic subordination of women. In common with other schools of heterodox economics, such as ecological economics, (old) institutionalism, and social economics, feminist economics is critical of the discipline's mainstream, also known as neoclassical, orthodox, free-market, or neoliberal

economics. Feminist economists, however, make the specific claim that orthodox economics naturalizes women's economic subordination and have also been critical of some other heterodox schools of thought, such as traditional Marxism, for the same reason.

A brief overview of some of the facts of economic inequality is in order.[2] In the United States, in 2007, full-time female workers earned 80 cents for every dollar earned by full-time male workers. In 2007, 12.5% of the U.S. population and 9.8% of all families lived in poverty. Single-parent households are more likely to be poor than two-parent households, and 85% of them are headed by women. Of those, 28.3% live in poverty. Of the 15% of single-parent households headed by males, only 13.6% live in poverty. Gender differentials in the global context are even more significant. Of the 1.3 billion people living in extreme poverty around the world, 70% are women. Globally, the average gender wage gap is 17%, but the range is from 3% to 51%. Two thirds of the world's working hours are performed by women, including the production of half of the world's food supply, but women receive only one tenth of the world's income and own 1% of the world's property.

However, globally and nationally, the differences between women can swamp the differences between men and women. For example, women in Botswana have a life expectancy of 33, but women in Hong Kong can expect to live until 86. When women work for pay in Georgia, they earn only 49 cents for every dollar earned by a Georgian man, while Maltese women earn 97 cents for every dollar earned by a Maltese man. In comparison to the earnings of full-time white male workers in the United States, full-time white female workers earn 79 cents for every dollar, full-time black female workers earn 68 cents, and full-time Hispanic female workers earn just 60 cents. Thus, many feminist economists are as concerned with the differences between women as they are with differences between men and women.

Feminist economists view these facts as problematic and symptomatic of an oppressive economic system that distributes economic rewards on the basis of gender, race, and ethnicity. They argue that mainstream economics plays an important role in the maintenance of this oppressive gender system through its various absences or silences around women and femininity, as well as race, class, and other signifiers of difference. Feminist economic research is both theoretical and empirical, but a piece of work cannot be called feminist economics if it does not support the feminist contentions that women are economically subordinate to men and that this is unacceptable. With this proviso in mind, one may exclude such classic economic tracts as Gary Becker's (1991) *Treatise on the Family,* which shows that, given a particular set of assumptions, it is efficient and hence, from a mainstream economic perspective, desirable for men to specialize in paid work and women in children and unpaid work. This understanding of the family is widespread among orthodox economists. Although there are many disagreements among feminist economists, all agree that Gary Becker's reductionist account of the sexual division of labor should not be included in any definition of feminist economics. Other mainstream economic treatments of women's economic status, even if sympathetic, also attribute most of the problem to women's choices, especially their innate desire to bear and rear children (a desire absent from men), and downplay the role of constraints on women's choices such as gender socialization and sex discrimination.

Given that feminist economists are united in their understanding of the current national and global distribution of economic rewards and punishments as gendered and as problematic, their primary aim is to improve the lives of women. But their research agendas can be very different, depending on the feminist perspective they take and their theoretical or empirical orientation. This plurality of approaches within feminist economics can, for convenience, be aggregated into two broad foci: research that uses gender as a theoretical variable and research that uses gender as an empirical variable. Those using gender as a theoretical variable are interested in the ways in which economic concepts and categories become gendered and in developing new theoretical perspectives on economic processes. Those feminist economists who undertake research in which gender is an empirical variable tend to be more practically minded, being interested in statistical analyses in which gender is a descriptive category and developing new statistical models to generate insights into the economic behavior of men and women.

## Gender as Theory and Practice

When a feminist economist uses gender as a theoretical variable or a category of analysis, she or he is taking a critical stance in relation to the power of the discipline to shape the way we see the world. Typically, she or he sees mainstream economics as a discourse, or a set of practices and institutions that do not merely describe or reflect a given reality but rather have an important role in creating and naturalizing the categories through which we interpret the world. Thus, mainstream economics is viewed as deeply implicated in the creation and reproduction of the hierarchies across gender, race, ethnicity, and other classifications of difference that privilege white heterosexual Western men at the expense of others. This kind of research problematizes and reconstructs orthodox and heterodox economic concepts and categories of analysis. It is crucial to the feminist economic project: After all, if feminist economics was an empirical endeavor alone, it would simply consist of economists who work on gender, with no critical feminist insight into the discipline's gendered foundations (see Barker, 2005; Hewitson, 1999).

As well as revealing the ways in which economic theories, frameworks of analysis, concepts, and categories are

gendered, feminist economists undertake empirical analyses that measure the real effects of theory, challenge existing empirical analyses justifying women's economic subordination, and offer new ways of examining theoretical issues. In these feminist economic works, gender is necessarily a descriptive rather than a theoretical variable—the aim of the work is to analyze and quantify the economic activities of actual men and women. Consider the following example of the two approaches. In the United States, women undertake most child care. Feminist economists using gender as a descriptive variable are interested in the impact of child care subsidies aimed at giving mothers access to the labor market on the same terms as those workers without child care responsibilities. Empirical analysis can provide estimates of the effects of the subsidy on mothers' labor supply behavior. A feminist economist might also develop an empirical model that includes variables not normally considered, such as state licensing requirements as a measure of quality. Furthermore, empirical feminist economists might advocate alternative ways of providing child care, such as mandated provision by employers (see Bergmann, 2005). In each case, the researcher is examining the economic activity of actual women, making gender an empirical category.

On the other hand, a researcher focusing on gender as a theoretical category might deconstruct the whole framework of analysis by examining the gendered meanings of the concepts of "mother" and "worker." The concept of mother has been historically and culturally constructed as someone who devotes herself to the well-being of her children. The concept of worker has been historically and culturally constructed as someone who devotes himself to a full-time job and a lifelong career, a breadwinner without domestic responsibilities that would reduce his work commitment. That is, mothers are women who care for children and do not work, while workers are male heads of households who undertake market work and not unpaid child care. This is not to say that men do not care for children or that mothers do not work for pay—it is an analysis of the concepts used in a debate about child care subsidies, and it would broaden that debate by questioning the institutions that reinforce the need for the "real worker" to be free of domestic responsibilities and therefore require that parenting be fitted into the workplace, rather than work fitted into parenting (see J. Williams, 2000). Both studies are vital: Empirical work provides essential information for policy decisions and brings theoretical work into the realm of economic practice, and theoretical work points to new ways of thinking about change and to new policies.

The following five sections review the key areas of theoretical and empirical research in feminist economics. Methodological concerns, reviewed in the first section, are crucial, as the methodology of a discipline dictates the authorized ways in which its practitioners go about producing the discipline's accepted knowledge. The second section deals with the rational economic agent, which is the foundational concept of mainstream economics, the concept upon which all its theories rely. Unpaid and paid work, those activities that structure most people's days, weeks, and years, are the subject of the third and fourth sections. In the final section, gender is examined within a global economic context.

## Methodology

Methodology refers to the way in which knowledge is generated and justified—the methods used by economists to establish what they know. Orthodox economists typically assert that economics is a science and that economic knowledge is produced using the scientific method. The scientific method refers to the process of statistically testing hypotheses against the facts in ways that can be replicated by any practitioner in the discipline. Knowledge develops when the facts support the hypothesis. Neoclassical economists view facts as independent of the hypotheses: They believe, therefore, that facts, or reality, can be perceived by anyone, whatever their social situation or value system. If this is so, it follows that the knowledge produced using the facts as the arbiter of truth must necessarily be objective. This is the basis of the claim that neoclassical economics is a "positive science" or knowledge from which normative statements or values have been excluded.

Feminist economists have been highly critical of these methodological claims. Along with other heterodox schools of thought, feminist economists reject the idea that neoclassical economics is value free and deny its claims to scientific status. Critics of neoclassical economics argue that it is underpinned by a conservative set of values based on a belief in individualism, the efficiency of free markets, and a merit-based system of economic rewards. This value system is built into the very fiber of the neoclassical paradigm: its methods, categories, and analyses. Feminist economists are unique in focusing on gender: They argue further that the value system underpinning neoclassical economics reflects the needs and preferences of masculinity as they have been constructed within the history of science.

Modern scientific methods emerged during the Enlightenment, when philosophers constructed a series of dualisms in their quest to understand how knowledge is produced. Rene Descartes, for example, with the dictum that "I think, therefore I am," constructed a mind–body split in which the thinker could be conceptually separated from any particular embodiment: This is the "view from nowhere" that defines the objective stance of the scientific inquirer. To see this, consider the meaning of an alternative, such as "I feel, therefore I am"—the thinker in this case is necessarily embodied, and embodiment is necessarily sexually and racially specific, which means that the "view" or perspective of the "feeling thinker" is from somewhere and hence not objective. Francis Bacon wrote about the necessity of conquering and penetrating nature,

thereby constructing a subject, the researcher, in opposition to the object of research, and in this can be seen the object–subject, mind–matter, and culture–nature distinctions. These distinctions were also gendered: The detached, objective observer with the "view from nowhere," the penetrator of passive matter, and the producer of scientific knowledge were associated with masculinity. Subjectivity, passivity, nature, the body, emotions, and materiality were associated, on the other hand, with femininity (see Nelson, 1993). These connections persist into the present day—one need only construct a list of contemporary stereotypical masculine and feminine characteristics.

The scientific method reflects a specifically *Western* as well as androcentric perspective. It was the European man, not just any man, who had the capacities of autonomy, independence, and objectivity that were necessary for scientific knowledge production. The scientific approach was becoming dominant over the seventeenth to nineteenth centuries, just when the West (or Europe) was discovering, conquering, and claiming as European territory the lands occupied by dark-skinned native peoples. Like nature, and women, these native peoples existed to be ruled by European men. They and their cultures were seen as passive, exploitable, primitive, and inferior, and science was used to justify their domination. The scientific method was used to establish the superiority of the colonizers and hence the unchallengeable supremacy of their perspective on the world. The power relations established by the use of the scientific method in economics continue to reverberate within economics. For example, in development economics, Western modernization via the extension of market relations is viewed as the solution to the lack of "progress" within the so-called less-developed world (see Olsen, 1994; R. M. Williams, 1993).

The dualisms and their gendered and racial associations created within this history of philosophical thinking define the way in which orthodox economists have understood the discipline's methodology—as scientific or positivist—from the late nineteenth century until today. Developments within the philosophy of science, particularly the rejection of the positivism espoused in every first-year economics textbook, have not undermined this self-perception. Feminist economists have drawn on feminist philosophy of science to argue that mainstream economists constitute a community of practitioners with a jointly held perspective and value system that underpins and structures the processes of knowledge production. Specifically, the perspective and value system is that of the middle-class, Western, white male who has historically dominated the discipline and who has been positioned by the scientific method as the source of unbiased knowledge. This value system is invisible within the mainstream, which is therefore incapable of recognizing its own gender and racial biases, and hence incapable of producing objective, or value-free, knowledge. But these biases powerfully shape, among the myriad of choices that are made by researchers in any discipline, the identification of research issues, the questions asked, the facts viewed as relevant, the methods deemed appropriate, the choice between taking as given and questioning the various assumptions that necessarily underlie the analysis of a problem, the interpretation of empirical results, and the choice of statistical methods to decide between hypothesis verification or falsification. The methodological argument, then, is not simply that the mainstream is dominated by men but that its method imposes a symbolically masculine perspective of the world on its practitioners, whether they are men or women (see the introduction and reprinted essays in Volume 4 of Barker & Kuiper, 2009; Strassmann & Polanyi, 1995).

For example, mainstream economists explain the gender wage gap as a function of human capital variables. That is, they assume that individuals' wages are the result of utility-maximizing choices with respect to their skills and qualifications. Competition ensures that labor markets equilibrate where the wage is equal to the marginal productivity of labor. Therefore, if women earn lower wages than men, mainstream economists' first response is that women must have lower productivity as the result of their utility-maximizing choices. The unquestioned assumptions underlying this analysis are numerous: that the individual is the appropriate unit of analysis, that productivity is an individual characteristic and not a function of the requirements of the job, that productivity is a choice variable, that preferences and technology are exogenous, that competitive market forces determine wages, that the marginal product of labor is observable independently of wages, that women expect to have and make decisions on the basis of a marginal connection to the labor market, that the problem must be amenable to mathematical modeling and statistical analysis, and so on.

A feminist economic analysis of the gender wage gap rejects the neoclassical view that wage setting is a rational and efficient process undertaken by individuals within competitive markets. Instead of viewing wage setting as an application of the mainstream's timeless and universally applied economic laws, wage setting is situated with a historically and culturally specific analysis where constructions of gender are central to understanding how work and workers are valued. Waged work emerged in the nineteenth century, unions formed, and men fought for a family wage, or a wage sufficient to sustain a family. It was in this era that the low values attributed to the service and caring occupations, within which women were crowded, were established, a function not of some allegedly objective standard like the marginal product of labor but of dominant social views on what women should be doing and how they should be doing it. In fact, men's demands for a family wage spells out these social views: Women should be in the home, economically dependent on a breadwinner. Tracing the impact of this history on contemporary valuations of women's work exposes the mainstream's value system—in ignoring gender constructions, in assuming the primacy of

the individual, in confining the analysis to the narrow limits of mathematical modeling, in viewing individuals as prior to society, and in assuming that preferences and technology are not shaped in relation to gender, among others. Mainstream economists can always justify sex wage differences as objectively determined differences in productivity, which are functions of individual preferences, because of what they carve off as irrelevant (see Figart, Mutari, & Power, 2002). An empirical feminist economic analysis challenges the neoclassical explanation within the data itself. The neoclassical story says that women expect to move in and out of the labor market because of their roles in the home and child rearing. It is therefore rational for them to choose to acquire human capital that retains its value during periods of absence from the labor market. It happens that these skills are also low-productivity ones, and hence, women earn less than men. Thus, we should expect to see different rates of depreciation of the education and skills required in male-dominated and female-dominated occupations. In fact, this hypothesis is not supported by the facts, but despite this, it remains a key plank in the neoclassical theory of wage differentials. Thus, the empirical approach, which adopts many of the methodological assumptions made by neoclassical economists, reveals the way in which the perspective of privileged white men shapes research agendas (see Blau, Ferber, & Winkler, 2010).

The mainstream's reliance on individual choices to explain social outcomes, such as the gender wage gap, occupational segregation, and women's poverty, is called methodological individualism. This method dictates that all analyses must be built on a foundation of the isolated individual, homo economicus. If society as a whole is simply the sum of isolated individuals, then the individual preexists that society, joining with other isolated individuals with a universal human nature and a fully formed set of preferences. Society has no role in creating individuals—as gendered and as raced—in this view; rather, society is a reflection of the universal human nature. In the next section, this human nature is discussed.

## The Rational Economic Agent

An important focus of feminist economists who use gender as a category of analysis has been the model of the individual at the center of all neoclassical theorizing. Several assumptions construct this model. The economic agent is assumed to embody a timeless human nature consisting of a number of critical characteristics: Individuals maximize their utility; they are instrumentally rational, always choosing the least-cost means to meet their objectives; they are self-interested and uninterested in the well-being of others; and they are independent or entirely separate from others. Socially significant categories such as race or ethnicity, class, gender, nationality, or sexuality are irrelevant to this definition of the economic agent. The characteristics attributed to the economic agent are critical because, without them, the neoclassical edifice of the modeling of individual choices within a constrained environment becomes hopelessly entangled. For example, in neoclassical consumer theory, rational, self-interested, and independent consumers maximize their individual well-being by spending their incomes such that the marginal utility of the last dollar spent on each good or service is equal. The solution to each consumer's maximization problem can be found diagrammatically as a point of tangency between the budget constraint and the indifference curve or mathematically by setting the derivatives of the utility function equal to zero. Now imagine the impact of a different theory of the economic agent: What if consumers care about others, live within complex social arrangements and relationships, and use ethics, rather than self-interest, to determine their shopping cart contents? This interrelatedness of individuals means that the solution to the consumer's utility-maximizing problem is a function of many other people's utility, potentially billions (e.g., when shopping fair trade). No neat diagrammatic or mathematical solutions are available: The predictions of individual behavior, as well as the whole policy framework of free markets that is built on those predictions, collapse.

Feminist economists have developed numerous critical analyses of the assumptions supporting the neoclassical theory of human nature (see England, 1993). As mentioned, neoclassical economics claims that this human nature is universal and preexists any social arrangements. This is the basis of the mainstream use of the literary figure of Robinson Crusoe as the representative economic agent. Crusoe was a British slave trader who was shipwrecked on a deserted island, on which he lived alone for more than two decades, and who then rescued a native, whom he called Friday, from cannibals. To neoclassical economists, the facts of Crusoe's race, sex, and class; his socialization in seventeenth-century London; the importance of the slave trade to the story; his unusual living conditions, including the absence of women, children, and a family; and his assertion of ownership of the island and virtual enslavement of Friday to his will are irrelevant to the capacity of the figure of Robinson Crusoe, a white Western colonizing man, to function as an exemplar of the economic agent. It need hardly be said that economic agents begin life as helpless babies and often end life as helpless elders. Economic agents then can, in reality, exist only within particular social and familial relations, relations that entail a fundamental dependence on others and are completely absent from the paradigmatic neoclassical story of the individual. Neoclassical economics excludes all these aspects of Crusoe's story, leaving only the fantasy of the autonomous agent, independent of all others, seeking only his own self-interest within competitive market conditions, naturalizing and legitimizing the failure of the mainstream to consider

gender, race, history, culture, power relations, and connections to others as absolutely essential to an understanding of the sexual, national, and global distribution of economic well-being (see Grapard & Hewitson, 2010).

## Unpaid Work

Unpaid work, including child care, shopping, subsistence crop production, food preparation, cleaning, laundry, and collecting water and firewood, is essential for the functioning of the market economy, by creating workers and consumers on a daily basis. Unpaid work absorbs as many hours of work as paid work, and the majority is performed by women. Several groups of feminist economists work in this area. Those who used gender as a theoretical category include Marxist feminist economists, who, in the 1970s, pointed out that the home was a site of production as well as consumption. How to integrate the idea that domestic labor was economic production into the existing concepts of the Marxist framework was the subject of the so-called domestic labor debate. Others sought to understand how unpaid work had come to be excluded from mainstream definitions of economic activity and the implications of this exclusion. Unpaid work as productive economic activity is also vital to the research agendas of empirical feminist economists working in areas such as national accounting, development, and labor markets (see the introduction and reprinted essays in Volume 2 of Barker & Kuiper, 2009).

Although unpaid work is an important aspect of the neoclassical economics of the family and its explanations for women's inferior economic outcomes, it really plays a minor role within the discipline as a whole, being almost or completely ignored in most research fields. This is a function of the way in which unpaid work evolved as a feminine-gendered concept and as an activity that takes place outside the realm of the economy per se. The foundational constrained optimization problem of labor economics, for example, is the utility-maximizing choice between paid work and leisure. And recent macroeconomic policy and performance debates are completely silent on unpaid work, despite the fact that the value of the output of home production rivals the value of market sector output (see Ironmonger, 1996). Outside of first-year economics classrooms, gross domestic product (GDP), which is the annual value of the market sector's output, is treated as a direct measure of the health and well-being of a nation. Although these exclusions seem natural to many economists, they rely not on some inevitable way of organizing economic activity but on a particular historically and culturally specific set of theoretical creations.

Before the Industrial Revolution (1770–1830), which heralded the widespread development of capitalism in Europe, the wage labor system did not exist. The economic unit was the family, and family members as well as any servants worked together to produce food, clothing, and perhaps some cash to buy things they could not make, such as tools. The family economy divided tasks by sex and age. For example, women might have been responsible for food preparation, milking, feeding livestock, and growing vegetables for home consumption; men for planting and harvesting grains (perhaps as a serf); and younger people for spinning and sewing. Task allocation varied by region, rather than being a set of male and female jobs common to all humanity. But the fact that people's work varied by sex and age did not imply a hierarchy of value. In the family economy, husbands and wives were equally essential for the survival of the family.

The spread of the capitalist mode of production, within which individuals sell their labor to employers and "go to work" at a central location, broke down the family economy and was the basis of the development of a hierarchical valuation of different types of work. A lot of the work undertaken by men left the house and attracted wage payments, while a lot of women's work did not. Work undertaken outside the home was often viewed as skilled work, while women's work was not—in fact, during the nineteenth century, women's work in the home lost its definition as work and became something that women did naturally and out of love for their families.

We see this transformation of unpaid work and the unpaid worker in the evolution of census categories during the nineteenth century. The censuses documented and categorized the population and its activities—in Britain, every 10 years from 1800. The categories used for this documentation were products of generally held views on gender, and men's and women's proper places, as well as the writings of economists. Early in the century, economists understood labor as the most important source of the wealth of nations; hence, the work that was undertaken by the population was of key significance. In the early decades of the census, those who worked in the home on domestic tasks were deemed to be economically occupied. Later in the century, however, economists excluded all nonmarket activities from their definition of economic activity, and by the end of the century, the census categories also reflected this new theoretical boundary of economic behavior. Thus, by the end of the century, women's work in British, Australian, and North American homes had no place within the census; rather, those undertaking domestic labor were categorized as economic dependents, or economically unoccupied (see Deacon, 1985; Folbre, 1991).

This particular history is responsible for many of the seemingly natural categories that are used to define and understand today's economies. For example, the labor force categories of employed, unemployed, and not in the labor force, as well as the national accounting system and GDP, are based on nineteenth-century census categories. Until 1993, the System of National Accounts (SNA), which generates estimates of the annual value of the productive activity in an economy, GDP, excluded unpaid

work (see Waring, 1990). The production boundary, or the division between productive and unproductive activities, enclosed the market and excluded nonmarket activities. In 1993, the SNA was revised, and the production boundary was extended to all goods for household consumption, whether or not those goods had been acquired through markets (United Nations Development Fund for Women [UNIFEM], n.d.). Where possible, the value of unpaid work is published in a satellite account. This means that macroeconomics, which is the study of GDP (its definition, how it changes, how its changes affect inflation and unemployment, and how the government can manage it), continues to exclude about half the economic activity actually being undertaken.

This is significant for many reasons. Without knowing anything about the household sector, economists cannot make claims about the efficiency of resource allocation. They are also blind to the full impact of economic policy (see Sharp, 1999). Cutting the federal budget, for example, may simply generate additional, invisible, unpaid work to be shouldered by women. Structural adjustment policies within the development context have been especially problematic in this regard. A full understanding of the amount and distribution of unpaid work is also vital to a full understanding of equity and welfare issues. It might not seem equitable, for example, that people working the same number of hours over a lifetime receive very different economic rewards, because in men's case, two thirds of their work is for pay, while in women's case, only one third of their work is paid. Furthermore, unpaid workers in the household sector, although imperative to the market economy, do not attract the benefits of paid work, such as social security. Nor are they covered by legislation that protects paid workers, such as occupational health and safety. Finally, without a full accounting of economic activity, all manner of distortions can persist. For example, income produced in the home in the form of goods and services is tax free, while income deriving from market activities is taxed.

Research into the value of unpaid work has taken two approaches. The first applies market wages to the hours of work in the home, while the second uses the value of the output produced in the home. The market wage method of valuing unpaid work can use three different market wages. Specialist wages can be used to value time spent on specialist tasks. For example, the time cooking a meal would attract a chef's hourly rate, and the time counseling children would receive a psychologist's hourly rate. Alternatively, a generalist wage can be used to value all the time spent in domestic labor, whatever the particular tasks. Here the value of unpaid work is the number of hours times the hourly wage of a housekeeper. Finally, the opportunity cost wage, or the wage that is given up to free the time for unpaid work, can be used to value unpaid work. These market wage applications normally generate an estimate of approximately half the value of GDP. Feminist economists have pointed out, however, that comparing the value of labor time in the home to GDP is like comparing apples to oranges. GDP is the sum of all incomes, not simply wages, and in particular, it includes the return to capital used up in the production process. Given the equality between GDP measured by incomes and GDP measured by the value of current production, comparing oranges to oranges requires that the value of the household sector be measured as the value of its output, or value added. When the economic activity of households is measured in this way, the value of the household sector is at least equal to the value of the market sector (see Beneria, 2003; Goldschmidt-Clermont, 1992; Ironmonger, 1996).

As noted, women undertake the majority of unpaid work. Feminist economists argue that the mainstream theory of this sexual division of labor personifies an ideal of the family and femininity that developed in post–World War II United States. With the growth of suburbia and the industrial war machine now manufacturing consumer goods, the 1950s saw the development of a powerful ideal of the modern suburban family with all the latest modcons, in which men had careers and women devoted their energies to keeping up with the Jones's, cleanliness, and helping their husbands' careers. It is this view of the family, reflected in television programs of the era such as *Leave It to Beaver,* which has been personified within the neoclassical theory of the family, primarily through the work of Gary Becker (1991). Becker's new home economics (NHE) modeled the household as a single unit, within which the wage earner is a benevolent dictator who seeks to maximize the household's well-being subject to constraints of time, wages, and prices. He can ensure, via distribution decisions, that each family member will concur with his wishes (maximize his utility). Benevolence guarantees that the household as a whole is as well off as it can be. Spouses exploit their comparative advantages, and hence husbands, rather than wives, will typically take on the role of benevolent dictator because men typically earn more than women. This is efficient because women can take advantage of economies of scale in childbearing and rearing—they can be pregnant while also caring for children.

Because NHE has been so influential and is the most widely used model of the household within the mainstream, feminist economists have attacked it vigorously (see Ferber, 2003). They have pointed out that the model relies on circular reasoning: Recall that husbands specialize in the labor market because they earn more, while wives specialize in domestic work because they earn less. But women earn less because they specialize in domestic work. To explain women's specialization in the home, look at women's lower wages. To explain women's lower wages, look at women's specialization in the home. This circular reasoning naturalizes women's role in the home and their lower labor market earnings by leaving out the possibility of labor market discrimination against women

and the role of gender ideologies and gendered institutions in shaping and forming value assessments of women's work and skills.

Another major problem with NHE's vision of the family is its silence on power relations. But empirical evidence suggests that power relations exist—in particular, that the person earning the most money wages has the most bargaining power. For example, the more equal are the wages of a husband and wife, the more equal is the division of unpaid labor. Those women who specialize in the home are likely to experience declines in their bargaining power over time as their labor market skills decline. Once considerations of power differentials enter the analysis, the notion that the sexual division of labor is efficient becomes extremely questionable (see Ferber, 2003). Instead of NHE, some feminist economists have used bargaining models to explicitly incorporate the power relations within families as well as including a longer term perspective than what is possible in the static NHE model (see the introduction and reprinted essays in Volume 2 of Barker & Kuiper, 2009).

A third important critique of NHE is its heteronormativity, or its assumption that natural family relations are heterosexual and reproductive. Indeed, Becker (1991) defines anyone who does not fit into such a family as "deviant." According to NHE, deviants are inefficient because they do not take advantage of the complementarity of men and women in reproduction and production. Such deviants include homosexual people, "career women," "house-husbands," people who do not want children, people who cannot have children, and people who prefer to remain single. This critique is also one internal to feminist economics because the category of the family is mostly taken by feminist economists themselves to be self-evidently made up of a heterosexual couple. The naturalization of the conjugal family contributes tremendously to economists' and policy makers' inability to imagine economic activity being organized and work, income, and wealth being distributed differently (see Badgett, 1995; Danby, 2007; Hewitson, 2003).

## Paid Work

Feminist economists argue that the gendered institutional structures that frame and reproduce the current organization of unpaid work also support women's economic subordination in the realm of paid work. Women's paid work often replicates their unpaid work, reflecting a gender ideology that maps femininity onto service work and work involving the support of men. Thus, women dominate in occupations such as maids and housekeeping cleaners, child care workers, elementary and primary school teachers, secretaries and administrative assistants, nurses, and receptionists. These jobs are both derivative of unpaid labor and often badly remunerated. Thus, the gender wage gap can also be traced to the sexual division of labor in paid work. NHE justifies this pattern of economic rewards but does not explain why, when women make up nearly half the labor force, they continue to be responsible for the majority of domestic labor and child care.

Interest in the interrelatedness of women's domestic role and their occupational distribution within the labor market has led to the development of a new category of analysis called *caring labor*. Caring labor refers to both paid and unpaid caring work, such as child care for pay and unpaid emotional support within the family. Feminist economists have found that caring occupations dominated by women tend to attract a "caring penalty," which can be linked to the lack of value attributed to unpaid work and the lack of esteem with which this work is generally viewed (see the introduction and reprinted essays in Volume 2 of Barker & Kuiper, 2009; Folbre, 1995).

Mainstream economists agree that there is a sexual division of labor in paid work and that there is a gender wage gap. However, as has been noted, they believe that these phenomena result from rational, utility-maximizing, individual choices. Early labor market studies within the orthodox school did not consider women at all. It was only in the 1960s, during an unprecedented movement of white wives into the formal labor market, that a female labor supply function was delineated, and because it referred to wives, it necessarily included the opportunity cost of women's time at work—not leisure but home-produced goods and services. Later, race and sex discrimination moved onto the mainstream agenda, though it was and continues to be argued that discrimination is an individual phenomenon ("I don't like black people or women"), analytically having nothing to do with larger social institutions such as the organization of unpaid work, gender ideologies, or the history of colonization, slavery, and associated racism. In the mainstream model, racists and sexists are punished by the market with lower profit than their competitors and hence go out of business.

There is an extensive empirical literature, which will not be reviewed in detail here, examining gender issues in the labor market from a feminist perspective (see Bergmann, 2005; Blau et al., 2010). Some key results will suffice. Feminist economists have found that sex discrimination plays a role in the gender wage gap. Women often earn less than men who are doing the same job and are promoted more slowly than equally or lesser qualified men. Women also hit a glass ceiling, so that in many occupations, the senior positions are largely taken by men, while women's careers have stopped progressing once they have reached some midway point up the ladder. Feminist economists also insist on the importance of "indirect discrimination," or the discriminatory impact of the gender system that shapes women's choices. When women choose to enter traditionally male occupations, they encounter the revolving door: Women enter, find the working environment hostile to women, and leave. Most occupations are

dominated by either women or men, and this occupational segregation also accounts for some of the gender wage gap. But as already discussed, the evidence for women's low productivity is lacking, and in any case, women can earn less even with the same human capital investments. Experiments have shown that application letters from male (or white) applicants are evaluated more positively and lead to more invitations to an interview than application letters that are the same in every relevant respect but are presented as being from female (or black) applicants.

Because of women's work in the home, feminist economists are very interested in the ways in which these responsibilities fit with the institutional requirements of the labor market, for full-time attendance at a workplace, a 40-hour week, 6 p.m. meetings, and so on. Because the categories of work and the worker are gendered masculine, as the complement of the feminine gendered unpaid work and the housewife (discussed above), the worker is someone without domestic responsibilities. (This is a theoretical point about the concept of the worker rather than being a point about empirical men and women.) This way of viewing the worker—that is, as a gendered category—leads to a different perspective on the labor market. Such mechanisms as the mommy track, family-friendly policies, and unpaid maternity and paternity leave are ways in which mothers are added to or fitted into the labor market, leaving mothers, rather than work, as the problem. Fathers, because the notion of worker is already intrinsically dependent on the idea of male breadwinner, have been hesitant to make use of these mechanisms for fear that their commitment to work will be questioned. Indeed, holding other factors constant, men with children earn more than those without, while the opposite is true of mothers, revealing the assumptions regarding the work commitment of breadwinners versus mothers. In other words, anyone not fitting the identity of worker in the same way as unencumbered men is problematic. This points once again to the extent to which gendered institutions naturalize and reproduce an organization of work that is detrimental to women (see Barker, 2005; J. Williams, 2000).

## Gender in a Global Perspective

In the early 1970s, Western feminists began to consider the role of women within Third World development. This literature came to be known as "women in development" (WID). It examined the ways in which Western modernization affected women's work, a topic neglected by earlier development specialists. Although the WID perspective added a much-needed voice to the development literature, which had thus far ignored the gendered impacts of development policies, it did so problematically. WID theorists viewed the women, men, and economies of "less-developed" countries (the global South) through Western concepts and categories. For example, they accepted that development meant the extension of markets and commodification but did not recognize that equality in labor market participation, being a goal of Western feminists, did not necessarily improve women's status or well-being in non-Western countries (see the introduction and reprinted essays in Volume 3 of Barker & Kuiper, 2009).

The fact that women of the South are not the same as Western women seems fairly obvious. But at a theoretical level, this insight is very powerful. It means that the category of woman is a construct of theory. In particular, it emerges from the privileging of the perspective of the West, discussed above. Western feminists are positioned within the history of European theorizing of knowledge production as colonizers, as superior, civilized, and modern, compared to women of the South. A stay-at-home white mother married to a wealthy man living in the United States, for example, has very different experiences of womanhood than a poor woman working in an Asian export processing zone, and an analysis of the desires, opportunities, and constraints facing the stay-at-home mother cannot be assumed to apply equally to the Asian woman. This is not simply because the Asian woman's situation is so different but because the very meaning of woman is different in each case. One is not just a woman, but a woman with a race. Feminist economists who generalize from the experiences of white Western women are performing an act of violence, in that it enacts an imperial power relation of colonizer and colonized and therefore silences non-Western women with different histories and cultures (see R. M. Williams, 1993).

The inability of the category of woman to be universally applied is part of the postcolonial critique that has developed within feminist economics. Feminist economists have also problematized, as gender and race specific, such seemingly natural entities and concepts as the national economy as an object of economic control, development, progress, and less-developed or underdeveloped countries. The concept of the national economy as it has been integrated into mainstream economics is a gendered and racial concept that developed during the mid-twentieth century at the time macroeconomics and the national accounting framework emerged from the Keynesian revolution and just as former colonies became independent. The new national economies, no longer subject to the needs of their colonizers, were understood to be in need of Western-style modernization, which marginalized the unpaid subsistence work of women (see Bergeron, 2004). The concepts of development and progress are situated firmly within the European theories of race and human evolution, which justified the creation of European empires and colonized peoples by positioning European societies as the endpoint of civilization (see Bergeron, 2004; Olsen, 1994; Zein-Elabdin & Charusheela, 2004).

There are also many important empirical issues in the area of gender in the global context. For example, the economies of the South have been the object of surveillance and control by the World Bank and the International Monetary Fund, and many have been subjected to structural adjustment policies, which reorient the economy to the

repayment of debt by minimizing the size of government and maximizing exports. But the full effects of structural adjustment programs were unknown because the subsistence and reproductive work mainly performed by women was not quantified or accounted for or integrated into the policies. In fact, however, increases in this work were critical for mediating the social costs of the policies, such as the loss of the safety net as governments cut their expenditures (see the introduction and reprinted essays in Volume 3 of Barker & Kuiper, 2009; Beneria, 2003).

The international sexual division of labor is another important area of research. Women make up the majority of workers in the factories of global corporations located in the South. They are paid poverty-level wages and lack basic protections such as occupational health and safety regulations and union membership. Global corporations will simply relocate should worker demands raise labor costs. Furthermore, hundreds of thousands of women from the South work as poorly paid domestics, maids, and nannies for wealthy women in the West and in wealthy Arab countries. Again we see the importance of the postcolonial critique and the emptiness of the assumption of some kind of global sisterhood (see Olsen, 1994).

## Conclusion

This chapter has established that the field of feminist economics is necessary, that feminist economic researchers use two main approaches, each with its own important role to play in meeting the goal of ending the economic subordination of women, and reviewed five of the most important areas within which these researchers are working. Needless to say, many other important areas of theory and empirical work could not be covered for reasons of space (see Peterson & Lewis, 1999). Examples include the history of economic thought (where Pujol [1992] uses gender as a theoretical variable, while Dimand, Dimand, & Forget [2001] take an empirical approach); feminist perspectives on other heterodox schools of thought, such as post-Keynesian economics and institutionalism; integrating feminist thinking into the teaching of economics (see Bartlett, 1997); engendering government budgets and macroeconomic policy (see Sharp, 1999); and gender in emerging market economies. The five topics discussed above, however, capture the essence of feminist economics: They cover the main ways in which the economic experiences of men and women are shaped by gender, and they address the most central and problematic of the concepts and categories of the mainstream.

## Notes

1. Feminist economics is quite recent. Its organization, the International Association for Feminist Economics, was formed in 1992, and its journal, *Feminist Economics,* began in 1995.

2. All statistics are from the United Nations Development Fund for Women (UNIFEM, n.d.) and the U.S. Census (n.d.).

## References and Further Readings


Badgett, M. V. L. (1995). Gender, sexuality, and sexual orientation: All in the feminist family? *Feminist Economics, 1*(1), 121–139.

Barker, D. K. (2005). Beyond women and economics: Rereading "women's work." *Signs, 30,* 2189–2209.

Barker, D. K., & Kuiper, E. (Eds.). (2003). *Toward a feminist philosophy of economics.* New York: Routledge.

Barker, D. K., & Kuiper, E. (2009). *Feminist economics* (4 vols.). New York: Routledge.

Bartlett, R. L. (Ed.). (1997). *Introducing race and gender into economics.* New York: Routledge.

Becker, G. S. (1991). *Treatise on the family* (Rev. ed.). Cambridge, MA: Harvard University Press.

Beneria, L. (2003). *Gender, development and globalization.* New York: Routledge.

Bergeron, S. (2004). *Fragments of development.* Ann Arbor: University of Michigan Press.

Bergmann, B. (2005). *The economic emergence of women* (2nd ed.). New York: Palgrave Macmillan.

Blau, F. D., Ferber, M. A., & Winkler, A. E. (2010). *The economics of women, men and work* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Danby, C. (2007). Political economy and the closet: Heteronormativity in feminist economics. *Feminist Economics, 13*(2), 29–54.

Deacon, D. (1985). Political arithmetic: The nineteenth century Australian census and the construction of the dependent woman. *Signs, 11,* 27–47.

Dimand, R., Dimand, M. A., & Forget, E. (2001). *A biographical dictionary of women economists.* Aldershot, UK: Edward Elgar.

England, P. (1993). The separative self: Androcentric bias in neoclassical assumptions. In M. A. Ferber & J. A. Nelson (Eds.), *Beyond economic man* (pp. 37–53). Chicago: University of Chicago Press.

Ferber, M. A. (2003). A feminist critique of the neoclassical theory of the family. In K. S. Moe (Ed.), *Women, family and work* (pp. 9–23). Oxford, UK: Blackwell.

Ferber, M. A., & Nelson, J. A. (Eds.). (1993). *Beyond economic man.* Chicago: University of Chicago Press.

Ferber, M. A., & Nelson, J. A. (Eds.). (2003). *Feminist economics today.* Chicago: University of Chicago Press.

Figart, D. M., Mutari, E., & Power, M. (2002). *Living wages, equal wages.* New York: Routledge.

Folbre, N. (1991). The unproductive housewife: Her evolution in nineteenth century economic thought. *Signs, 16,* 463–484.

Folbre, N. (1995). "Holding hands at midnight": The paradox of caring labor. *Feminist Economics, 1*(1), 73–92.

Goldschmidt-Clermont, L. (1992). Measuring households' nonmonetary production. In P. Ekins & M. Max-Neef (Eds.), *Real-life economics* (pp. 265–283). New York: Routledge.

Grapard, U., & Hewitson, G. J. (Eds.). (2010). *Robinson Crusoe's economic man.* New York: Routledge.

Hewitson, G. J. (1999). *Feminist economics.* Cheltenham, UK: Edward Elgar.

Hewitson, G. J. (2003). Domestic labor and gender identity: Are all women carers? In D. K. Barker & E. Kuiper (Eds.), *Toward a feminist philosophy of economics* (pp. 266–283). New York: Routledge.

Ironmonger, D. (1996). Counting outputs, capital inputs and caring labor: Estimating gross household product. *Feminist Economics, 2*(3), 37–64.

Kuiper, E., & Sap, J. (Eds.). (1995). *Out of the margin.* New York: Routledge.

Nelson, J. A. (1993). The study of choice or the study of provisioning? Gender and the definition of economics. In M. A. Ferber & J. A. Nelson (Eds.), *Beyond economic man* (pp. 23–36). Chicago: University of Chicago Press.

Nelson, J. A. (1996). *Feminism, objectivity and economics.* New York: Routledge.

Olsen, P. (1994). Feminism and science reconsidered: Insights from the margin. In J. Peterson & D. Brown (Eds.), *The economic status of women under capitalism* (pp. 77–94). Aldershot, UK: Edward Elgar.

Peterson, J., & Lewis, M. (Eds.). (1999). *The Elgar companion to feminist economics.* Cheltenham, UK: Edward Elgar.

Pujol, M. A. (1992). *Feminism and anti-feminism in early economic thought.* Aldershot, UK: Edward Elgar.

Seguino, S., Stevens, T., & Lutz, M. A. (1996). Gender and cooperative behavior: Economic *man* rides alone. *Feminist Economics, 2*(1), 1–21.

Sharp, R. (1999). Women's budgets. In J. Peterson & M. Lewis (Eds.), *The Elgar companion to feminist economics* (pp. 764–770). Cheltenham, UK: Edward Elgar.

Strassmann, D., & Polanyi, L. (1995). The economist as storyteller: What the texts reveal. In E. Kuiper & J. Sap (Eds.), *Out of the margin* (pp. 129–150). New York: Routledge.

United Nations Development Fund for Women (UNIFEM). (n.d.). United Nations development fund for women. Available at http://www.unifem.org

U.S. Census. (n.d.) United States Census [Online]. Available at http://www.census.gov

Waring, M. (1990). *If women counted.* New York: HarperCollins.

Williams, J. (2000). *Unbending gender.* Oxford, UK: Oxford University Press.

Williams, R. M. (1993). Race, deconstruction, and the emergent agenda of feminist economic theory. In M. A. Ferber & J. A. Nelson (Eds.), *Beyond economic man* (pp. 144–153). Chicago: University of Chicago Press.

Zein-Elabdin, E., & Charusheela, S. (Eds.). (2004). *Postcolonialism meets economics.* New York: Routledge.

# 89

# Neuroeconomics

Dante Monique Pirouz
*University of Western Ontario*

Decision making is a fundamental part of human behavior. We all make decisions every day that influence our health, well-being, finances, and future prospects, among other things. Researchers have become increasingly interested in why we make the decisions we do, especially when, in many cases, these decisions do not appear to be rational or beneficial to us in the long run. While neoclassical economics has traditionally looked at how people *should* behave, other disciplines such as psychology and cognitive science have tried to answer the question of *why* people act the way they do.

A new discipline, referred to as neuroeconomics, has sought to meld theory and methodology from diverse areas such as economics, psychology, neuroscience, and decision theory to create a model of human behavior that not only explains but also predicts how people make decisions (Glimcher & Rustichini, 2004). Neuroeconomics research examines how people make choices and attempts to determine the underlying neural basis for these choices and decisions. This chapter examines some of the seminal studies in neuroeconomics, highlighting the public policy implications and offering areas of future research where neuroeconomics could be applied.

## Theory

Traditional economic theory has maintained that humans are rational decision-making entities, that each individual has a clear sense of his or her own preferences, tries to maximize his or her own well-being, and makes consistent choices over time (Huang, 2005). However, this model is more often violated than upheld as people and animals attempt to outwit evolution and destiny. Neuroscience gives researchers the opportunity to look into the "black box" of cognitive processing to reveal empirical indications of how the brain really processes choice, risk, and preferences. The goal is to create "a complete neuroeconomic theory of the brain" (Glimcher, 2003).

Decision theory integrates mathematics and statistics to better understand how decisions such as choices between incommensurable commodities, choice under uncertainty, intertemporal choice, and social choice are made. It has been assumed that agents respond rationally in forming their choices and preferences. This theory finds that any "normal" preference relation over a finite set of states can be expressed as an expected utility equation.

However, the introduction of prospect theory, which suggests the possibility that other factors may affect behavioral decision making for the individual, has generated an interest in understanding the underlying mechanisms of preference, judgment, and choice (Kahneman & Tversky, 1979). The significance of these findings can have important implications for the marketing discipline. To this end, a better understanding of the decision-making processes used by people is important to understanding the critical drivers of economic behavior.

Psychology has sought to investigate the inner workings of the human mind (Camerer, Loewenstein, & Prelec, 2005; Loewenstein, Rick, & Cohen, 2008). Cognitive psychology, and more recently cognitive neuroscience, has introduced new tools that allow researchers to capture and measure data from brain activity related to a specific function and behavior. This new type of data has led to new directions of research that combine neuroscience, psychology, and decision theory to better understand the complexities of human decision making.

Neuroscience looks at the structure, function, and development of the nervous system and brain, while cognitive neuroscience investigates how behavior and the nervous system work together in humans and animals. In other words, cognitive neuroscience is the study of the neural mechanisms of cognition (Gazzaniga, 2002). At the nexus of neuroscience, economics, and psychology, there is an area that has tentatively been coined *neuroeconomics,* which uses neuroscience techniques to look specifically at how human subjects make choices. Neuroeconomics is interested not only in exposing brain regions associated with specific behavior but also in identifying neural circuits or systems of specialized regions that control choice, preference, and judgment (Camerer et al., 2005; Loewenstein et al., 2008).

Techniques borrowed from neuroscience include brain imaging methods that may reveal how humans and animals use the neural substrates of the brain to process and evaluate decisions, weigh risk and reward, and learn to trust others in transactions. Brain imaging techniques that can be used on human subjects include electroencephalography (EEG), positron emission tomography (PET), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI). EEG and MEG measure changes in electrical brain currents by placing electrodes on the scalp to measure electrical waves emitted from the cortex. PET scans measure changes in blood flow by capturing images of radioactive isotopes injected in the bloodstream. fMRI measures blood flow to neural regions by relying on the magnetic properties of oxygenated and deoxygenated blood in the brain.

Another technique borrowed from neuroscience is transcranial magnetic stimulation (TMS), which is used to produce a magnetic pulse that can temporarily interfere with normal brain activity. For example, TMS can produce sudden movements in motor areas. While it has not yet been used for neuroeconomic studies, TMS has been used successfully for cognitive neuroscience studies and could potentially be used in the future to study decision making. In addition, there are invasive techniques of monitoring brain activity in animals, including single-cell recording, wherein an electrode is passed through the skull into the brain and neural activity is recorded. Neuroeconomics studies can also use nonimaging techniques. For example, some studies have been conducted using patients with brain lesions that disable specific parts of the brain. In addition, to determine central nervous system (CNS) response, studies can measure hormone levels, galvanic skin response, sweat gland activity, and heart rate (Carter & Tiffany, 1999; Frackowiak et al., 2004).

Before these technologies made it possible to examine the neural mechanisms of cognition, much of economic theory relied on the rational choice model. This model posits that individuals have stable preferences and a clear understanding of the options facing them. Thus, people are assumed to make their choices based on careful, unemotional calculations that maximize the benefits and minimize the costs that they will incur. However, current models of decision making only partially explain real human behavior. Neuroeconomics examines higher level cognitive functions of personal choice and decision making, demonstrating how these are expressed at the neuronal and biochemical levels. The analysis of this newest form of data that more closely examines brain processing promises to bring us closer to answering questions as to why people consume, have addictions, save, and hoard; what drives preference and choice; and what makes people happy, risk seeking or risk adverse, and trusting or trustworthy.

Over the past 50 years, scientists have experimented with a number of hypothetical game scenarios to determine models of how people make choices in economic situations. Before imaging technology, it was not possible to accurately investigate the influence of emotions and cognition on these economic models of decision making. However, behavioral economists have begun to challenge the assumptions of the rational agent and have found that psychological and emotional factors do indeed play an important role in people's economic decision-making process. Essentially, neuroeconomics looks at two branches of choice: solitary choice and strategic choice (Zak, 2004).

In their article on neuroeconomics, Camerer et al. (2005) argue for the fundamental insights that neuroscience could offer economics. They maintain that economic theory has assumed that agents can "mentalize," or infer from the actions of others, what their preferences and beliefs are. However, accumulating evidence from individuals with autism, Asperger's syndrome, and brain lesions shows that mentalizing is a specialized skill modularized in specific brain regions. More important, the ability to mentalize exists in varying degrees from person to person.

## Applications and Empirical Evidence

### The Neuroscience of Game Playing

Games give neuroeconomists a useful way to isolate decision and choice variables in experimental studies. Most of these studies look at either behavior, autonomic reactions (such as hormone levels or heart rate), or brain activity while subjects are engaged in strategic games, thus revealing how the neuronal system processes fairness, reward, loss, trust, distrust, revenge, discounting, and choice. Specific brain regions have been implicated in how judgments are made about perceptual stimuli received from our environment (Adolphs, 2003). Some of these brain regions involved in judgment include the amygdala, which is central in the processing and memory of emotions; the insula, believed to be involved with feelings of disgust and unease; and the anterior cingulate cortex, which is implicated in reward anticipation, decision making, and empathy (Frackowiak et al., 2004).

## Neural Calculations of Decisions

The idea that people seem prone to violate expected utility theory has led to the development of alternative models on how choices are made under risk. One such alternative, prospect theory, exhibits a series of effects that alter the value assigned to gains and losses (Kahneman & Tversky, 1979). Phenomena such as the certainty effect, which states that people are prone to undervalue probable versus certain outcomes, or the isolation effect, which finds inconsistent preferences for identical outcomes based on how the outcomes are framed, challenge the notion that utility theory holds in real-life cases of human judgment (Kahneman & Tversky, 1982). Interestingly, Camerer et al. (2005) make clear in their article on neuroeconomics that all the violations of the utility theory that humans commit have been replicated in animal studies. For example, rats have also committed the same patterned violations in addition to other expected utility properties (Kagel, Battalio, & Green, 1995). Probability is how animals and humans calculate associations between events and predict outcomes critical to survival and understanding their environment. For example, there is evidence that dopamine neurons of the primate ventral midbrain may act to predict reward by specifically coding errors (Fiorillo, Tobler, & Schultz, 2003). It was found that dopamine levels increase during gambling, which indicates that uncertainty may be the mechanism that induces this dopamine rush. This may explain the reward people feel when gambling, which cannot be explained by the monetary gain of gambling because losses usually outnumber gains.

## Trust and Cooperation

Imaging studies have revealed more about how social interaction shapes neural response, allowing us to choose mutual cooperation and shared gains over self-interested choices to create a sense of stability in longer-term game scenarios (McCabe, Houser, Ryan, Smith, & Trouard, 2001). Increased activity in players who were more trusting and cooperative was shown in a brain area believed to be the locus of mentalizing, as well as in the limbic system, where emotions are believed to be processed (Camerer et al., 2005).

The trust game is often used in neuroeconomics studies and mimics the relationship between an investor and broker. The game is played in multiple rounds where a player is given an amount of money (e.g., $10) and then must decide how much of the money, if any, to send to a second player. The amount is tripled, and then the second player decides how much to send back to the first player. One study found that subjects who received the first "investment" violated the rational response, which would be to accept any amount of money offered to them by the first player or "investor" (McCabe, 2003). Interestingly, when a small amount of money was offered by a computer player instead of a human, the response by the investor was not as extreme. In other words, players were upset about receiving a low return only if they believed that another person was trying to take advantage of them. If they thought the small amount of money was from an impartial computer, investors were not as emotionally sensitive. In addition, half of the subjects in the study were characterized as cooperators and had a common pattern of divergent activation in the prefrontal cortex where simultaneous attention to mutual gains and inhibition of immediate gratification allow for cooperative choices.

Studies of fMRI brain scans found that when responders were offered low monetary amounts (e.g., $1 out of a maximum of $10), there was more activation in the prefrontal cortex, anterior cingulate, and the insula cortex (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). When monetary offers were low, the receivers had increased activity in the insula, which is often associated with feelings of pain and disgust (Wright, He, Shapira, Goodman, & Liu, 2004). The anterior cingulate cortex receives input from a number of other areas and is thought to resolve conflict among these areas. A player refusing an offer could be predicted by the level of activation in the insula. The author speculates that the insula may be the neural area responsible for distaste for inequality or unfair conditions.

The relationship between trust and hormones is another area of interest for researchers. Hormonal response in people was investigated during a series of trust games to determine whether there was a specific hormone that could be connected to feelings of trust and distrust (Zak, Kurzban, & Matzner, 2004). Participants' blood was tested after each round for levels of oxytocin, which has been associated in facilitating social behaviors, social recognition, maternal attachment, pair bonding, and the feeling of falling in love. The study found that when money was returned to the first player, oxytocin did indeed increase to twice the levels of the random draw. This means that if people felt they were being trusted, increased oxytocin levels made them more likely to trust back. Interestingly, ovulating women were less likely than nonovulating women or men to give money back even if they received the full amount from the other player. This, Zak (2004) believes, is due to the fact that progesterone, which increases during ovulation, binds with oxytocin to inhibit its affect. In looking at distrust, Zak looked at dihydrotestosterone and testosterone in both men and women to see if levels increased during low-trust games. The study found that testosterone did not significantly increase in either women or men, and dihydrotestosterone levels did not increase for women. However, there was a significant increase in the level of dihydrotestosterone in men when the other player signaled distrust. Zak hypothesizes that this may be related to the increased feelings of aggression that men reported when engaged in a low-trust game.

While cooperation is an important component in human society, the desire to punish is the flip side of cooperation, which may be how society is able to enforce social norms.

An interesting study that looked at the neural basis for altruistic punishment or revenge found that people feel a sense of satisfaction when punishing those who break what are considered social norms (de Quervain et al., 2004). Using PET scans, researchers found greater activation in the striatum, which is usually "implicated in the processing of rewards that accrue as a result of goal-directed actions." In addition, those with the strongest responses in the striatum were more likely to take on higher costs for the right to mete out punishment to those who deviate from societal norms.

## Fairness

Humans tend to reject inequality even if it means walking away from a reward (Powell, 2003). In a study that looked at the neural substrates of cognitive and emotional processing, specifically fairness and unfairness activated during the ultimatum game, 19 subjects were scanned using fMRI (Sanfey et al., 2003). The ultimatum game is based on one player offering the other a split of a sum of money that the responder can either reject or accept. Players were paired with others who offered various split amounts of $10. The responders were scanned as they decided whether they would choose fair or unfair proposals. Previous behavioral research on the ultimatum game found that low offers are rejected 50% of the time even though a rational maximizing solution would be for the responder to accept any amount of money because some money should be better than no money. Subjects usually report that low offers are often rejected because it provokes an angry response. In Sanfey et al.'s (2003) study, brain imaging revealed that unfair offers activated the bilateral anterior insula, dorsolateral prefrontal cortex, and anterior cingulate cortex. The anterior insula is often implicated in negative emotional responses, more specifically in disgust (Krolak-Salmon et al., 2002; Wright et al., 2004). The dorsolateral prefrontal cortex is often implicated in executive function and goal maintenance, which may stem from the responder actively maintaining the cognitive goal of acquiring as much money as possible. Increased activation of the insula was biased toward rejection of the offer, and increased activation of the dorsolateral prefrontal cortex was biased to accepting the offer. The anterior cingulate cortex has been implicated in cognitive conflict and may be a result of conflict between emotional and goal motivation during the game. Interestingly, the experiment was also run with both human and computer partners who acted in offering the split. The response in these brain areas was stronger when unfair offers were made by the human partner versus the computer, suggesting that the response was not just to the monetary amount offered but also to the contextual factor that the unfair offer was made by another human.

Even monkeys seem to adhere to this notion of fairness. Brosnan and de Waal (2003) found that cooperation may have developed through evolution where individuals must compare their own efforts to the payoff they receive with those of others. Brown capuchin monkeys responded negatively when offered unequal rewards from experimenters and even refused to participate when they witnessed other monkeys receiving more attractive rewards for the same amount of effort. The researchers posit that this inequity aversion may have an evolutionary origin in our neurological development.

## Reward and Loss

Kahneman and Tversky (1979) found that loss is judged by people as being more painful than an equivalent gain is pleasurable, as is evidenced in the convex utility curve for losses and concave utility curve for gains in the value function. How valuation of gain and loss is calculated in the brain is an area under investigation by neuroscientists. Montague and Berns (2002) have looked at a number of experiments to develop a computational model referred to as the predictor valuation model, which anticipates neural responses in the orbitofrontal cortex and striatum. Other brain imaging studies have found that the brain processes gains and losses differently (K. Smith, Dickhaut, McCabe, & Pardo, 2002). PET imaging has revealed that there are two separate but functionally integrated choice systems, both in anatomical structure and in processing, each sensitive to loss. The dorsomedial system processes loss when deliberating risky gambles. When subjects make a choice that results in loss, there is a greater use of the dorsomedial system, which serves to calculate the loss versus the visceral representations in the more primitive ventromedial system, which animals most likely use to make decisions. Choice processing seems to be centered in the more medial structures, with more ventral than dorsal distribution.

Animal studies, mainly using monkeys, are revealing new information about how animals estimate the value of specific actions. For example, in a series of experiments, Schultz (1998; Schultz, Dayan, & Montague, 1997) looked at the neuronal response in the substantia nigra and the ventral tegmental area of the monkey brain to determine activity when a monkey pressed levers for juice rewards. Another animal study looked at whether specific neuronal activation can be correlated to the probability that the animal expects gain (Platt & Glimcher, 1999). Another animal study looking at reward valuation has found that reward valuation in monkeys can be predicted in a model based on reward history that duplicates foraging behavior (Sugrue, Corrago, & Newsome, 2004).

How we respond to monetary reward has also been investigated using fMRI. The neural substrates of financial reinforcement overlap with areas that deal with primary reinforcers, such as food (Elliott, Newman, Longe, & Deakin, 2003). Gold (2003) has made an argument for reward expectation to be linked to the basal ganglia. Breiter, Aharon, Kahneman, Dale, and Schizgal (2001)

have used fMRI to analyze the neural response to expectation and experience of gains and loss. The study found that that there may be a common circuitry of neurons that processes different types of rewards. In studies conducted with monkeys, it was found that the rhinal cortex was important for creating the associations between visual stimuli and their motivational significance (Liu, Murray, & Richmond, 2000; Liu & Richmond, 2000). Monkeys whose rhinal cortex had been removed were not able to adjust their motivation to changes in a reward schedule, while unaffected monkeys were able to adjust their motivation. The complexity of how motivation works to cause action is not clearly understood, but it is believed that a limbic-striatal-pallidal circuit forms the basis for the translation of motivation into action (Liu et al., 2000).

The neural substrate association with time discounting was investigated using fMRI, and it was found that human subjects use different regions in the brain to calculate short- and long-term monetary rewards (McClure, Laibson, Loewenstein, & Cohen, 2004). The limbic system associated with dopamine production tended to be activated when decisions that would bring immediate gratification were contemplated. On the other hand, the lateral prefrontal cortex and posterior parietal cortex were activated regardless of whether there was short or long intertemporal delay. There was greater frontal parietal cortex activity only when the choice made by subjects was longer term.

## Risk

Aversion to risk is linked to the amygdala and is driven by the ancient fear response (Camerer et al., 2005). Cortical override of the fear response is demonstrated in animal studies using shock. Over time, the response will be "extinguished." However, when the connections between the amygdala and the cortex are severed in the animal, there is a tendency for the fear response to return. This demonstrates that the amygdala does not "forget" but that the cortex is suppressing the response.

Risky choice is different from risk judgment in that the subject must choose between risky gambles that force an interaction between cognition and affect. Patients with damage to the ventromedial prefrontal cortex seem to suffer from decision-making deficits. In a study that measured performance in gambling tasks, patients continued to make the wrong choice, resulting in higher losses even after knowing the correct strategy (Bechara, Damasio, Damasio, & Tranel, 1997). Normal subjects used the advantageous strategy even before they consciously realized which strategy worked best. In addition, normal subjects developed skin conductance responses when facing a risky choice even before they knew that the choice was actually risky. For the patients with prefrontal damage, these skin conductance responses never developed. This suggests that there might be a nonconscious, autonomic bias that guides risky decision making based in the ventromedial prefrontal cortex, responding even before conscious cognition is aware of the risk. Bechara et al. (1997) hypothesize that this covert bias activation is dependent on past reward and loss experiences and the emotions that go with them, with damage to the ventromedial cortex interfering with access to this knowledge.

A study looked at how emotion affects perceptions of risk in investment behavior. Using patients with lesions in the brain areas associated with emotion, Shiv, Loewenstein, Bechara, Damasio, and Damasio (2005) compared investment decisions over 20 rounds to those made by patients with lesions in areas unrelated to emotion (control) and normal subjects. They hypothesized that the patients with damage to the emotional regions would be able to make better investment decisions because they would not be subject to emotional reactions that could lead to poor choices. This hypothesis was based on the case of a patient with ventromedial prefrontal damage who was able to avoid an accident on an icy road while others skidded out of control. The patient revealed later that because he felt a lack of fear, he was able to calmly react to the road conditions by thinking rationally about the appropriate driving response (Damasio, 1994). This led the researchers to wonder whether a lack of normal emotional reactions might allow people to make more advantageous decisions.

The study found that normal participants and control patients became more conservative in the investment strategy after a win or loss, whereas the lesion patients took more risk and, as a result, made more money from their investing choices (Shiv et al., 2005). Other studies have found that even low levels of negative emotions can result in loss of self-control, which can have less than optimal outcomes for the subjects. For example, as the result of myopic loss aversion, people exhibit high levels of loss aversion when gambles are presented one at a time rather than all at once (Benartzi & Thaler, 1995).

## Addictive Behavior

Closely related to the question of reward is addiction. Neural activity drives the search for food in both animals and humans. These same neuronal networks may also drive behavior to seek other kinds of substances that rate high on the reward evaluation. When the brain is strongly activated by, for example, sugar, food, or drugs, it can lead to abuse, which is often called addiction (Hoebel, Rada, Mark, & Pothos, 1999). A critical issue in behavioral decision theory is the question of why people and animals would choose to engage in behavior that is detrimental or harmful. This issue is related to the question of addictive behavior and the endeavor to understand the neural underpinnings of reinforcement and inhibition of behavior. In an early neuroeconomic paper dealing with this issue, Bernheim and Rangel (2002a, 2002b) proposed a mathematical theory of addiction that sought to explain irrational addictive behavior in

terms of decision theory and economics. The model is based on the idea that cognitive processes such as attention can affect behavioral outcomes regardless of initial preference. If a person is subject to "hot cognition" (or affect-laden thinking), for example, he or she may engage in consumption behavior that conflicts with preference because the focus is on usage and "the high" (Bernheim & Rangel, 2002a).

The theory of cue reactivity is another theory that might serve to explain why addiction levels remain high even though subjects self-report that they are striving to quit and they do not enjoy the consumption of their addictive substance (Carter & Tiffany, 1999; Laibson, 2001). In a recent study that used fMRI to analyze neural response in adolescents to alcohol-related imagery, researchers found that adolescents with even a short usage history of alcohol had significantly higher blood oxygen response in areas of the brain associated with reward, affect, and recall (Tapert et al., 2003).

Various types of addictive behaviors are under investigation by researchers. A type of consumption addiction involves the dispensation of products. The neural basis of collecting and hoarding in humans was analyzed using patients with prefrontal cortex lesions (Anderson, Damasio, & Damasio, 2005). In addition to judgments of use and consequences of discarding possessions, other cognitive processing going on at the time of the decision to save may be important. Compulsive hoarders appear to have a peculiar perspective with regard to possessions. When deciding whether to discard a possession, they spend most of their time thinking about being without the possession (the cost of discarding) and little time thinking about the cost of saving it or the benefit of not having it. This notion is similar to an observation made by J. P. Smith (1990) about animal hoarding. Smith speculated that the sight of a nut (by a squirrel) puts the squirrel "in touch with" the feeling of being hungry and without the nut. For the hoarders, the sight of the possession puts them "in touch with" the feeling of being without the possession and needing it. This feeling dominates their consideration of whether the possession should be discarded (Frost & Hartl, 1996).

## Limitations

The fact that brain science offers insights into economic and behavioral phenomena is not necessarily a new concept. While not universally embraced by all economists, some behavioral economists have been using constructs from psychology to attempt to build more descriptive and realistic models of behavior (Huang, 2005). However, for the first time, imaging technology such as fMRI offers the type of tools that can effectively explore the subtleties of the human brain while being noninvasive, relatively safe for human subjects, and providing results that are robust and revealing. However, fMRI studies have been questioned by critics because of, for example, use of small sample sizes (typically less than 40 subjects), ambiguity in human neuroanatomy mapping, lag time in the hemodynamic response, image distortion due to signal drop-off, motion artifacts, poor temporal resolution, and the debate over functional definitions of neural areas (Savoy, 1998, 1999; Savoy, Ravicz, & Gollub, 2000; Wald, 2005). Despite the limitations and difficulties in analyzing the results produced by fMRI, significant improvements in brain mapping, imaging power, and resolution (there are now 3T, 4T, and 7 T scanners being used to gain improved imaging resolution) have indicated that at least some of these shortcoming may be reduced with the next generation of equipment.

## Policy Implications

An example of how neuroeconomics could be applied to an important research area deals with the question of consumption addiction. This is especially true in developing effective marketing communications for vulnerable consumers such as children and adolescents. Pechmann, Levine, Loughlin, and Leslie (2005) presented evidence from the addiction and neuroscience literature that adolescents were more vulnerable to advertising and promotions due to the unique structure of their neural development. This may indicate that the decision-making process for adolescents is significantly different from that of children and adults. While there has been evidence from empirical social psychology studies to support this assertion, increasing evidence based on neuroimaging studies has been developed by researchers from psychology, neuroscience, and medicine (Pechmann & Pirouz, 2007). This important development could offer a strong basis for research that seeks to investigate how and why adolescents respond to marketing communications and advertising differently. Furthermore, research could begin to develop ways to protect vulnerable adolescents from detrimental product categories such as cigarettes and alcohol, while enhancing the relevance and efficacy of marketing, such as health messages, targeted toward adolescents. In this way, neuroeconomics methods can offer researchers a valuable suite of methods that will allow a more refined and revealing understanding of the neural basis of choice for adolescents—a developmentally unique segment of the population.

## Future Directions

A diverse array of questions can be addressed using neuroeconomic techniques and methods. Neuroeconomics could serve as an important new area for tackling many of the fundamental questions about decision making that have been difficult to explain theoretically. Neuroeconomics offers the potential for insights into the neurological processes that

underlie human behavior. Using experimental methodologies combined with imaging and other neuroscience tools, neuroeconomics can better help us understand the mechanisms of decision making, including preference, risk behavior, valuation, biases, and conflict. While neuroeconomics as a field of study is in a relatively early phase, a growing number of researchers are establishing new theoretical constructs that could potentially inform economics, behavioral decision theory, management, marketing, and psychology.

Within neuroeconomics, a number of intriguing areas of research have not yet been fully explored and could prove of further interest. Such future areas of research might include the following:

- How do neural systems work together to create decision-making behavior?
- How does decision making vary for vulnerable populations such as adolescents or the elderly?
- What factors influence the development of addictive behavior, and what factors could act to discontinue these addictions?
- How can an understanding of the neural systems underlying decision making help people to make better decisions in their lives?

## Conclusion

While the application of neuroscientific methods to economics and other related fields may cause continuing controversy and debate among scientists and the public, the results gleaned thus far from neuroeconomic research have revealed valuable insights into the neural substrates that affect human and animal decision making. It seems reasonable to think that these insights may allow for new, more revealing models of decision making that will take into account the underlying neurological mechanisms that drive behavior, emotion, and choice.

## References and Further Readings


Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. *Neuroscience, 4,* 165–178.

Anderson, S. W., Damasio, A. R., & Damasio, H. (2005). A neural basis for collecting behaviour in humans. *Brain, 128,* 201–212.

Bechara, A., Damasio, A. R., Damasio, H., & Tranel, D. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275,* 1293–1295.

Benartzi, S., & Thaler, R. (1995). Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics, 110,* 73–92.

Bernheim, B. D., & Rangel, A. (2002a). *Addiction and cue-conditioned cognitive processes.* Cambridge, MA: National Bureau of Economic Research Program on Public Economics.

Bernheim, B. D., & Rangel, A. (2002b). *Addiction, cognition and the visceral brain.* Cambridge, MA: National Bureau of Economic Research Program on Public Economics.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A. M., & Schizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron, 30,* 619–639.

Brosnan, S. F., & de Waal, F. B. M. (2003). Monkeys reject unequal pay. *Nature, 425,* 297–299.

Camerer, C. F., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature, 43,* 9–64.

Carter, B. L., & Tiffany, S. T. (1999). Meta-analysis of cue-reactivity in addiction research. *Addiction, 94,* 327–340.

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain.* New York: HarperCollins.

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science, 305,* 1254–1258.

Elliott, R., Newman, J. L., Longe, O. A., & Deakin, J. F. W. (2003). Differential response patterns in the striatum and orbitofrontal cortex to financial reward in humans: A parametric functional magnetic resonance imaging study. *Journal of Neuroscience, 23,* 303–307.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299,* 1898–1902.

Frackowiak, R. S. J., Friston, K. J., Frith, C. D., Dolan, R. J., Price, C. J., Zeki, S., et al. (2004). *Human brain function* (2nd ed.). San Diego: Elsevier Academic Press.

Frost, R. O., & Hartl, T. L. (1996). A cognitive-behavioral model of compulsive hoarding. *Behaviour Research and Therapy, 34,* 341–350.

Gazzaniga, M. (2002). *Cognitive neuroscience.* New York: W. W. Norton.

Glimcher, P. W. (2003). *Decisions, uncertainty and the brain.* Cambridge: MIT Press.

Glimcher, P. W., & Rustichini, A. (2004). Neuroeconomics: The consilience of brain and decision. *Science, 306,* 447–452.

Gold, J. I. (2003). Linking reward expectation to behavior in the basal ganglia. *Trends in Neurosciences, 26*(1), 12.

Hoebel, B. G., Rada, P., Mark, G. P., & Pothos, E. N. (1999). Neural systems for reinforcement and inhibition of behavior: Relevance to eating, addiction and depression. In D. Kahneman & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology.* New York: Russell Sage.

Huang, G. T. (2005, May). The economics of brains. *Technology Review,* p. 74.

Kagel, J. H., Battalio, R., & Green, L. (1995). *Economic choice theory: An experimental analysis of animal behavior.* New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–292.

Kahneman, D., & Tversky, A. (1982). Judgement under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Krolak-Salmon, P., Henaff, M.-A., Isnard, J., Tallon-Baudry, C., Geunot, M., Vighetto, A., et al. (2002). An attention modulated response to disgust in human ventral anterior insula. *Annals of Neurology, 53,* 446–453.

Laibson, D. I. (2001). A cue theory of consumption. *Quarterly Journal of Economics, 116,* 81–120.

Liu, Z., Murray, E. A., & Richmond, B. J. (2000). Learning motivational significance of visual cues for reward schedules requires rhinal cortex. *Nature Neuroscience, 3,* 1307–1315.

Liu, Z., & Richmond, B. J. (2000). Response differences in monkey TE and perirhinal cortex: Stimulus association related to reward schedules. *Journal of Neurophysiology, 83,* 1677–1692.

Loewenstein, G., Rick, S., & Cohen, J. D. (2008). Neuroeconomics. *Annual Review of Psychology, 59,* 647–672.

McCabe, K. (2003). A cognitive theory of reciprocal exchange. In E. Ostrom & J. Walker (Eds.), *Trust and reciprocity: Interdisciplinary lessons from experimental psychology.* New York: Russell Sage.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences USA, 98,* 11832–11835.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science, 306,* 503–507.

Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron, 36,* 265–284.

Pechmann, C., Levine, L. J., Loughlin, S., & Leslie, F. (2005, Fall). Self-conscious and impulsive: Adolescents' vulnerability to advertising and promotions. *Journal of Public Policy & Marketing,* pp. 202–221.

Pechmann, C., & Pirouz, D. (2007, June). *The dark side of attachment: Addiction.* Paper presented at Advertising and Consumer Psychology 2007: New Frontiers in Branding: Attitudes, Attachments, and Relationships, Santa Monica, CA.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature, 400,* 233–238.

Powell, K. (2003). Economy of the mind. *PLoS Biology, 1,* 312–315.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300,* 1755–1758.

Savoy, R. (1998). *Introduction to designing fMRI-based experiments.* Boston: Massachusetts General Hospital, Martinos Center for Biomedical Imaging.

Savoy, R. (1999). Magnetic resonance imaging (MRI). In G. Adelman & B. H. Smith (Eds.), *Encyclopedia of neuroscience* (2nd ed.). Boston: Elsevier.

Savoy, R., Ravicz, M. E., & Gollub, R. (2000). The psychophysiological laboratory in the magnet: Stimulus delivery, response recording, and safety. In C. T. W. Moonen & P. A. Bandettini (Eds.), *Functional MRI.* Berlin: Springer.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80,* 1–27.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599.

Shiv, B., Loewenstein, G., Bechara, A., Damasio, A., & Damasio, H. (2005). Investment behavior and the negative side of emotion. *Psychological Science, 16,* 435–439.

Smith, J. P. (1990). *Mammalian behavior: The theory and the science.* Tuckahoe, NY: Bench Mark Books.

Smith, K., Dickhaut, J., McCabe, K., & Pardo, J. V. (2002). Neuronal substrates for choice under ambiguity, risk, gains, and losses. *Management Science, 48,* 711–718.

Sugrue, L. P., Corrago, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science, 304,* 1782–1787.

Tapert, S. F., Cheung, E. H., Brown, G. G., Lawrence, F. R., Paulus, M. P., Schweinsburg, A. D., et al. (2003). Neural response to alcohol stimuli in adolescents with alcohol use disorder. *Archives of General Psychiatry, 60,* 727–735.

Wald, L. (2005). Physics of NMR, MRI, functional MRI, and safety. In *Functional MRI visiting fellowship: A short course in fMRI.* Boston: Massachusetts General Hospital, Martinos Center for Biomedical Imaging.

Wright, P., He, G., Shapira, N. A., Goodman, W. K., & Liu, Y. (2004). Disgust and the insula: fMRI responses to pictures of mutilation and contamination. *NeuroReport, 15,* 2347–2351.

Zak, P. J. (2004). Neuroeconomics. *Philosophical Transactions of The Royal Society, 359,* 1737–1748.

Zak, P. J., Borja, K., Matzner, W. T., & Kurzban, R. (2005). The neuroeconomics of distrust: Sex differences in behavior and physiology. *Cognitive Neuroscience Foundations of Behavior, 95,* 360–363.

Zak, P. J., Kurzban, R., & Matzner, W. T. (2004). The neurobiology of trust. *Annals of the New York Academy of Sciences, 1032,* 224–227.

# 90

# Evolutionary Economics

Clifford S. Poirot Jr.
*Shawnee State University*

Evolutionary economics has gained increasing acceptance as a field of economics that focuses on change over time in the process of material provisioning (production, distribution, and consumption) and the social institutions that surround that process. It is closely related to, and often draws on research in, other disciplines such as economic sociology, economic anthropology, and international political economy. It has important implications for many other fields in economics, including, but not limited to, growth theory, economic development, economic history, political economy, history of thought, gender economics, industrial organization, the study of business cycles, and financial crises.

Historically, evolutionary economics was the province of critics of the mainstream, neoclassical tradition. Both Marxist and original institutional economists (OIE) have long asserted the importance and relevance of understanding change over time and critiqued the standard competitive model for its abstract, ahistorical, and static focus. In recent years, however, the rise of the new institutional economics (NIE) as well as game theory has resulted in wider acceptance of evolutionary explanations by the mainstream (Hodgson, 2007b, pp. 1–15; North, 1990). Consequently, it is now possible to identify three major traditions in evolutionary economics: the Marxist (Sherman, 2006), the OIE (Hodgson, 2004), and the NIE (North, 1990). Each of these major traditions encompasses multiple strands within it. As a general rule, Marxists and OIEs seek to replace the standard competitive model of mainstream economics, while NIEs seek to complement the standard competitive model, although the growing acceptance of game theory may make this less of an important distinction. Despite their differences, it is possible to identify some common themes that are shared by each of these disparate traditions. For example, authors in each tradition have exhibited a concern with how the interaction of technology, social institutions, and ideologies leads to changes in economic and social organization over time.

The goal of this chapter is to introduce the reader to a few of the major concerns, themes, and important authors of each respective tradition. In doing so, it will first address some general issues in evolutionary economics, including its relationship to evolutionary biology as well as some conceptual, definitional, and taxonomic issues. It will then proceed to provide a brief overview of the evolution of each respective tradition. Unfortunately, the length of this entry precludes discussion of many worthy contributions to each tradition as well as important topics that can and should be addressed by evolutionary economics. For example, space does not permit a discussion of how evolutionary economics could be applied to gender economics or how economists who write on gender often incorporate the contributions of evolutionary economists. Nor will this entry attempt to assess the extent of empirical or conceptual progress in evolutionary economics within or between respective traditions. In addition, the reader should be aware that evolutionary economics itself is an evolving field and that the boundaries between the three traditions are often fluid.

## General Issues

### Relationship Between Theories of Biological and Sociocultural Evolution

Taken at face value, the word *evolution* simply means change. But Darwin's theory of gradual (step-by-step) evolution by variation of inherited characteristics and natural selection (differential survival based on the level of adaptation)

removed both theological and teleological explanations from the process of biological evolution and placed humans firmly in the natural world. The modern neo-Darwinian synthetic theory of evolution combines Darwin's focus on gradual (step-by-step) change based on variation of inherited characteristics and natural selection with modern population genetics. Both Darwin's original theory and the modern synthetic theory of evolution explain change within a species, the rise of new species, and the more dramatic kinds of change such as the rise of mammals, primates, and eventually human beings as a result of the same step-by-step process (Mayr, 2001, 2004).

At the risk of oversimplifying slightly, it should be noted that the neo-Darwinian synthesis formulated by Thedosius Dobzhansky and Ernst Mayr in the 1950s has given rise to two sometimes opposing strands within the overarching frame of the synthesis (Mayr, 2004, pp. 133–138). One strand, exemplified by Richard Dawkins, who has written many widely read books on evolution, focuses on the role of genes in building organisms and on the tendency of natural selection to result in highly adapted organisms. This approach is sometimes referred to as the strong adaptationist program in evolutionary biology. It is closely related to fields such as sociobiology and evolutionary psychology, which explain many human behaviors in terms of their evolutionary origins.

Other evolutionary biologists have de-emphasized the role of natural selection and emphasize the importance of understanding biological evolution in terms of emergence, chance, path dependence, satisficing, and punctuated equilibrium. Richard Lewontin and the late Stephen J. Gould are two widely read authors who have advocated this position. Both Gould and Lewontin have been strongly critical of biologically based explanations for human behavior.

Although these two differing approaches to evolution are sometime viewed as rivals, they are in actuality complementary to each other. It is important to understand both aspects of biological evolution. In addition, biological evolution is a very complex process, and evolutionary biologists continue to push their field forward. Contemporary research in evolutionary biology focuses on the important interactions between genes, organisms, and their interaction with the environment in the process of development. Evolutionary biologists have also become more aware of the importance of lateral gene transfer and endo-symbiosis in bacteria evolution. However, there is still widespread consensus among evolutionary biologists that the synthetic theory of evolution is a true theory. Evolutionary biologists reject theories that incorporate teleological explanations or inheritance of acquired characteristics because these theories have been discredited empirically. Evolutionary biologists reject theories that are premised on or seek to find evidence of supernatural design as this adds nothing to the explanation and draws the focus of science away from understanding and explaining natural law.

Evolutionary economists often draw on and incorporate concepts developed by evolutionary biologists to explain how economic evolution occurs. For example, many evolutionary economists view economic evolution as a nondirected step-by-step process that is non-teleological (it lacks a specific goal or predetermined endpoint). Many, although not necessarily all, evolutionary economists agree that humans have at least some genetically based cognitive and social predispositions that are a result of genetic evolution. Some examples include the ability to learn a language, to learn social norms, to cooperate in groups, and to develop complex tool kits with which to transform nature into useable goods and services. In addition, the use of the Darwinian concepts of inheritance, variation, and selection as analogs to explain outcomes is pervasive in evolutionary economics. Evolutionary economists also distinguish between specific or microevolution (change that occurs within a sociocultural system) and general or macroevolution (change from one sociocultural system to another).

Some evolutionary economists view the market as natural and as an extended phenotype. Other evolutionary economists argue that evolutionary economics should be viewed as a generalization of the Darwinian concepts of variation, inheritance, and natural selection with each case specifying additional, relevant detail (Hodgson, 2007a; Hodgson & Knudsen, 2006). Others have argued that while Darwinian concepts often provide useful analogies for understanding sociocultural evolution, aspects of sociocultural evolution are distinctly non-Darwinian (Poirot, 2007). For example, in at least some instances, social and economic evolution results from the conscious decisions of groups of purposive agents who intentionally design or redesign human institutions. Also, in the process of sociocultural evolution, we can pass on cultural traits that we acquire through the process of learning. Biological evolution results in a branching pattern and barriers between different species. But human cultures can always learn from each other. The more emphasis that is placed on purposive design of social institutions and cultural learning as well as the abruptness (instead of the step-by-step nature) of social change, the less Darwinian a model of sociocultural evolution becomes. However, it would be difficult to identify anyone today who argued for a strong teleological concept of sociocultural evolution or who sought to explain sociocultural evolution in terms of divine or supernatural intervention.

Two other important concepts borrowed from the natural sciences, emergence and complexity, also play a key role in evolutionary economics. Emergence means that an observed system results from the complex interaction of the components of the subsystems. This process of interaction gives rise to patterns that would not be predicted from and cannot be reduced to the behaviors of the individual components. However, understanding the system still requires an understanding of its components and the

interaction of the components. So it is important to understand what individuals do. And it is also important to understand how individual choices and habits interact with social institutions in a dynamic way. It is often easier to think in mechanical terms. But if we are careless with mechanical analogies, then we can be easily misled.

This raises the question of what it is that evolves in sociocultural evolution. In evolutionary biology, selection takes place at multiple levels but logically requires changes in the gene pool of a population over time (Mayr, 2004, pp. 133–158). This has led some evolutionary economists to suggest that institutions and/or organizational routines provide us with an analog to the gene. Others argue that there is not a precise analog. To understand this debate, we first have to understand what an institution is.

It is popular to define institutions as "rules of the game." This is a good start, but it confuses the function of institutions with a definition of institutions. A more extensive definition of *institution* defines an institution as any instituted process, or in other words a shared, learned, ordered, patterned, and ongoing way of thinking, feeling, and acting. Institutions may be tacit and informal or highly organized and structured. By this latter definition, modern firms, medieval manors, technology, nation-states, political ideologies, and even technology are all institutions. In other words, virtually everything that humans do is an instituted process. Institutions are component parts of a sociocultural system.

But to just call everything an "institution" can make it difficult to conduct analysis. So it is useful to draw a distinction between entities such as social ideologies (e.g., Calvinism and democracy), social institutions (e.g., class, caste, kinship, the family, the nation-state), organizations (e.g., the modern firm, the International Monetary Fund, the medieval manor), organizational routines of actors within specific organizations, and technology (the combined set of knowledge, practices, and tool kits used in production). So in that sense, everything in sociocultural systems is constantly evolving. There is no precise analog in sociocultural evolution to the gene pool of a population.

As suggested above, social institutions are part of more general wholes, which it is convenient to term *sociocultural systems.* A sociocultural system includes the direct patterns of interaction of a society with the ecosystem (its subsistence strategy, technology, and demographic patterns), its social institutions, and its patterns of abstract meaning and value. Many anthropologists classify sociocultural systems by their scale, complexity, and the amount of energy captured by their subsistence strategy. Standard classification includes bands, tribes, chiefdoms, agrarian states, and industrial states, each of which corresponds roughly to subsistence strategies of foraging, horticulture, pastoralism and fishing, settled agriculture, and modern industrial technology. This classification system provides a useful scheme with which to understand the rise of large agrarian empires in the neolithic era and, ultimately, the Industrial Revolution in northwestern Europe. It also provides a useful classificatory schema with which to understand the interaction of multiple kinds of contemporary societies in a globalizing world. However, care must be taken to emphasize the multilinear and dynamic nature of sociocultural evolution rather than rigidly applying these concepts as a universal and unilinear schema (e.g., see Harris, 1997; Wolf, 1982).

### The Scope and Methods of Evolutionary Economics

The evolutionary biologist Ernst Mayr (2004) argued that biologists who study genetic evolution ask "why" questions while biologists who study things such as biochemistry ask "how" questions. Similarly, many mainstream economists ask "how" questions while evolutionary economists ask "why" questions. While the study of evolutionary economics does not preclude the use of formal mathematical models or quantification, most of its practitioners employ qualitative and interpretive methods. Also, as suggested above, some evolutionary biologists focus on changes that occur at the level of species, while others focus on more dramatic kinds of change. Similarly, evolutionary economists are interested in the study of sociocultural evolution on a grand scale, such as the rise of agrarian empires or modern capitalism, as well more specific, micro-level evolution such as changes in the organizational routines of individual firms.

Consequently, the kinds of issues that evolutionary economists are interested in overlap with the focus of other social sciences and even, in some instances, with the fields of ecology and evolutionary biology. Evolutionary economics reflects a tendency to counter the fragmentation of political economy into disparate social sciences that occurred in the late nineteenth and early twentieth centuries. Evolutionary economists, like their counterparts in economic sociology, economic anthropology, and political economy, focus more directly on those institutions with the strongest, most immediate, direct relevance to the process of material provisioning. So there may still be a need for some division of labor in the social sciences. What is of direct relevance will vary according to what is being analyzed in any particular study. An economic historian studying the rise of capitalism may, following Weber, find an understanding of Calvinist theology to be essential. Someone studying financial innovation in twenty-first-century industrialized societies would most likely find the religious affiliation of modern banking executives to be of little interest or relevance.

## Research Traditions in Evolutionary Economics

Evolutionary economics is composed of three rival but sometimes overlapping major traditions: the Marxist, the OIE, and the NIE. While there is some degree of ideological

overlap between the schools, each of the respective schools tends to share a common overarching ideology. Marxists seek to replace capitalism, OIEs seek to reform capitalism, and NIEs generally view capitalism as beneficent. This is not, notably, to argue that the ideology necessarily *determines* the empirical and theoretical analysis. Also, as previously noted, Marxists and OIEs seek to replace the standard competitive model while NIEs seek to complement the standard model. However, the reader should be aware that the boundary between the three traditions is often fuzzy, and there is sometimes overlap between the three traditions. Similarly, each of these three schools is composed of multiple strands and has undergone significant change over time.

The remainder of this entry will focus on outlining in very broad terms a few of the significant themes and concerns of each respective tradition, how these traditions have changed over time, and the contributions of a few representative authors of each of the three traditions. The reader may note that despite the differences between the traditions, there is a strong interest in all three in understanding how technology, social institutions, and cognitive models interact in the process of sociocultural evolution. The division made between the three traditions may be of greater interest and relevance in the United States, where there is a strong correlation between specific organizations and schools of thought. For example, the Association for Evolutionary Economics (AFEE) has been the primary promoter of OIE in the United States. In contrast, the European Association for Evolutionary Political Economy (EAEPE) has a much wider umbrella. So there may be hope someday for a grand synthesis of the three respective traditions.

## Marxist Models of Evolution

There are, of course, many different Marxist and quasi-Marxist models of sociocultural evolution. For the purposes of this entry, it is convenient to make the *differentia specifica* of a Marxist model of sociocultural evolution a focus on class struggle: the conflict between social groups defined in terms of differential access to the productive resources of a given society (Dugger & Sherman, 2000). This way of understanding sociocultural evolution is often referred to as historical materialism. While Darwinian reasoning may at times be employed in Marxist theories of sociocultural evolution, Marxists have generally emphasized the non-Darwinian aspects of sociocultural evolution as well as sharp discontinuities between human and infrahuman species. At the same time, it is hard to think of any academic Marxists writing today who would advocate Lysenkoism or Lamarckian theories of inheritance as valid explanatory concepts for understanding genetic evolution.

To understand historical materialism, we must begin with Marx's concept of the mode of production (for extended discussions, see Wolf, 1982, chap. 3, and also Fusfeld, 1977). A mode of production includes the techno-environmental relationships (e.g., agriculture based on a plough or factories using steam engines) and the social relationships of production (e.g., warlords and peasants or factory owners and workers) or, in Marxist jargon, the forces of production and the social relations of production, respectively. These relationships between groups of people in Marx's view are characterized by unequal relations of power, domination, subordination, and exploitation. This gives rise to social conflict over the terms of access to and the distribution of the productive resources of society. Social conflict requires the creation of a coercive entity to enforce the interests of the dominant social class (i.e., a state). In addition, human beings develop complex ideologies with which to justify their positions. Thus, the entire civilization (or what above is termed a *sociocultural system*) rests on a given mode of production, with the mode of production distinguished by the primary means of mobilizing labor (e.g., slavery, serfdom, wage labor).

In his analysis of Western history, Marx distinguished between the primitive commune, the slave mode of production of the ancient Roman Empire, the Germanic mode of production, the feudal mode of production of medieval Europe, and the modern capitalist mode of production. In analyzing Western history, Marx argued that each successive mode of production had produced technological advance, thus elevating the material level of human existence.

Capitalism, in Marx's view, is qualitatively different from extended commodity production. Capitalism requires that land, labor, and capital are fully treated as commodities. This means that labor is "free" in the sense of not being legally bound to perform labor for the dominant class and "free" in the sense that it has no claim to the resources needed to produce goods and services. Therefore, capital is used as a means to finance innovation in production, and labor is compelled by economic circumstances to sell its labor power. Because capitalism promotes endless accumulation of capital, it is thus far the most successful in a material sense. However, the dynamic of capitalist accumulation gives rise to periodic crises, and it is therefore unstable. In addition, it is often destructive of human relationships. So a relationship of apparent freedom is in actuality a relationship of power, subordination, and domination that will give rise to social conflict. The only way to end this conflict, in Marx's view, is to redesign social institutions so as to promote both development of the forces of production and social cooperation (i.e., replace capitalism with socialism). There is disagreement among scholars who study Marx as to whether Marx thought that the triumph of socialism over capitalism was inevitable.

Insofar as one seeks to explain the historical origins of capitalism and the Industrial Revolution, two historical epochs are of particular relevance. Marxist historians and Marxist economists (and many others) with a particular interest in economic history thus often refer to two transitions (one from antiquity to feudalism and the other from feudalism to capitalism) as giving rise to modern capitalism. Howard Sherman (1995, 2006), a well-known Marxist economist, has summarized and synthesized much of this existing literature.

Sherman traces Western economic history from tribal organization through the rise of modern capitalism. Sherman is a materialist who analyzes societies by starting with the material base of human existence and examines the interaction between technology, economic institutions, social institutions, and ideologies. Technology and technological innovation as well as social conflict between classes are key variables in Sherman's analysis. But overall, Sherman's schema is holistic and interactive, rather than mechanical or reductionist.

In analyzing the breakdown of feudalism, Sherman focuses on the tripartite class conflict between peasants, nobles, and monarchs and the ability of each of the respective classes to force an outcome on the other classes. As a consequence of this conflict, a new pattern of relationships based on private property and production for profit in a market, as well as increasingly organized around new sources of mechanical power, gave rise to a unique and extremely productive system referred to as capitalism. This system of production encourages constant cost cutting, innovation, and capital accumulation, thus leading to the potential for the progressive material elevation of human society.

However, capitalist society is still riven by conflict between property-less workers and property-owning capitalists. Because the capitalist has a monopoly over the productive resources of society, the capitalist is still able to compel the worker to produce a surplus for the capitalist. This creates social conflict between the capitalist and worker and also forces the capitalist into an ultimately self-defeating boom-and-bust cycle of rising profits and increasing concentrations of capital, followed by falling rates of profit, leading to cycles of recession and crisis. The institutional structure of capitalism also magnifies other social conflicts and problems such as environmental degradation and destruction, as well as relations between racial and ethnic groups and genders. The solution to this social conflict, in Sherman's view, is to replace the institutions of capitalism with economic democracy (i.e., democratic socialism).

Sherman, who has long been a critic of Stalinist-style socialism, also extends his analysis to change in Russia and the Soviet Union. The October Revolution of 1917 occurred because neither the czar nor the Mensheviks were able to satisfy the material aspirations of the vast majority of Russians. But industrialization in the Soviet Union became a nondemocratic, elite-directed process due primarily to the particular circumstances surrounding the Bolshevik Revolution, the ensuing civil war, and the problems of the New Economic Policy. In time, factions among the elites developed as the Soviet economy proved unable to satisfy the material aspirations of the majority of the Soviet population. This created new pressure for change as elites were able to capture this process. Due also to pressure from the West, change in the former Soviet Union took the direction of restoring capitalism rather than developing greater economic democracy.

It should be noted that the standard Marxist model of historical materialism focuses on the ability of capitalism to elevate the material *capacity* of human societies. This focus has been challenged by the rise of world systems and dependency theory. Theorists who follow this line of thinking focus on the uneven nature of development and the tendency of core economies to place boundaries on the development of formerly colonized areas of the world. Some theorists in this tradition have been justly accused of having a rather muddled conception of the term *capitalism,* insofar as they claim inspiration from Marx. The late Eric Wolf (1982), a well-known economic anthropologist, resolved many of these conceptual issues in his book *Europe and the People Without History.* So rather than assume that capitalism leads uniformly to material progress, Wolf extended the historical materialist model to analyze the process of uneven development in the world system as a whole. In their textbook on economic development, James Cypher and James Dietz (2004) provide an excellent history and exposition of classical Marxism, dependency theory, and extended analysis and discussion of the new institutional economics, original institutional economics, and modernization theory.

## Original Institutional Economics

Thorstein Veblen (1898) was the founder of OIE, and his influence on OIE continues to be prevalent (Hodgson, 2004). Veblen was strongly influenced by Darwin's theory of biological evolution and held evolutionary science as the standard for the social sciences, including economics, to emulate. He was also deeply influenced by the evolutionary epistemology of the American pragmatists Charles Saunders Peirce and John Dewey. In addition, he incorporated the contrasting positions of nineteenth-century evolutionist anthropology, as exhibited by the work of Tylor and Morgan, and the historical particularism of Franz Boas. Although he was strongly critical of Marx and of Marxism, there are both parallels as well as differences in the writings of Marx and Veblen.

Like Marx, Veblen focused on the importance of understanding the interaction of changes in technology, social institutions, and social ideologies as well as social conflict. Veblen also had a stage theory of history, which he borrowed from the prevailing anthropological schemas of his day. However, where Marx focuses on concepts such as class and mode of production, Veblen focuses on instituted processes and the conflicts created by vested interests seeking to reinforce invidious distinctions. Veblen's model of sociocultural evolution is a conflict model in that it focuses broadly on social conflict that arises in the struggle for access to power, prestige, and property. But it is not a class-based model in the sense that Marxists use class.

In "Why Is Economics Not an Evolutionary Science?" (1898) and in "The Preconceptions of Economic Science" (1899), Veblen developed a critique of the mainstream economics of his day. In developing this critique, Veblen was critical of the abstract and a priori nature of much of

mainstream economic analysis. In articulating this point, he contrasted the "a priori method" with the "matter of fact method." This particular aspect of Veblen's criticism has often led some to view both Veblen and later OIEs as "atheoretical." But this misses the point for at least two reasons.

Veblen did not eschew theoretical analysis per se. He was however, critical of theory that divorced itself from understanding actual, real-world processes of material provisioning. But most important, in Veblen's view, economics was not up to the standards of evolutionary science because economics continued to implicitly embrace the concepts of natural price and natural law by focusing on economics as the study of economizing behavior and the adjustment of markets to equilibrium. In contrast, Veblen argued that the process of material provisioning entailed a constant process of adaptation to the physical and social environment through the adjustment of institutions or deeply ingrained social habits based on instinct. Veblen's understanding of the term *institution* was broad enough to encompass any instituted process. Yet he drew a sharp distinction between institutions and technology. He was sharply critical of the former and strongly in favor of the latter.

When Veblen wrote about deep-seated and persistent social habits developing on the basis of genetically based instincts, he did in fact appear to mean something similar to contemporary theories of gene-culture evolution. Social habits are not consciously thought-through, purposive behaviors—they develop out of the complex "reflex arc" of enculturation based on genetically based propensities to act in the presence of environmental stimuli. Instincts are acquired through genetic evolution and social habits through enculturation. Both are inherited, vary in nature, and may therefore be selected for or against in the process of sociocultural evolution (Hodgson, 2004, Part III). However, Veblen also borrowed from Dewey a view of socialization in which individuals are active participants in socialization, a concept that was later more clearly articulated by Meade. In addition, Veblen also emphasized the ability of humans to conceptualize and engage in purposive behavior.

Veblen drew a sharp dichotomy between the instinct of workmanship and the instinct of predation. He associated the instinct of workmanship with a focus on adaptive, problem-solving, tinkering, and innovative behavior. In contrast, he associated predation with a focus on brute force, ceremonial displays of power, emulative behavior, conspicuous consumption, financial speculation, and the power of vested interests. Veblen argued that the instinct of workmanship arose in the primitive stage of human history (roughly corresponding to what contemporary anthropologists would term *bands* and *tribes*) and that the instinct of predation emerged during the stage of barbarism (roughly corresponding to the rise of chiefdoms). These instincts gave rise to deep-seated social habits. Both instincts continued to be present during the rise of civilization (agrarian states) and persisted in modern civilization (industrial states). But because modern civilization is based on the rise and extensive application of machine technology, further progress would require the triumph of the instinct of workmanship over the instinct of predation.

But in Veblen's view, there was no reason to expect this would necessarily occur. Vested interests were often capable of instituting their power to reinforce the instinct of predation. Hence, institutions often served to encapsulate and reinforce the instinct of predation. The behaviors of predation were primarily exhibited by the new "leisure class" or, in other words, the robber barons of the late nineteenth century. In contrast, workmen and engineers often exhibited the instinct of workmanship. Consequently, Veblen tended to view institutions in general as change inhibiting and the instinct of workmanship as change promoting.

In later works, Veblen extended this kind of analysis to study other topics such as changes in firm organization and the business cycle. Veblen argued that as modern firms became larger and more monopolistic, a permanent leisure class arose, thus displacing technological thinking among this new class. In addition, increasing amounts of time and energy were channeled into financial speculation, leading to repeated financial crises. Emulative behavior in the form of conspicuous consumption and ceremonial displays of patriotism and militarism served to reinforce the instinct of predation. In his analysis of the rise of militarism in Prussia, Veblen noted the socially devastating impact of the triumph of the instinct of predation. Thus, Veblen tended to identify institutions with imbecilic behaviors that serve to block the triumph of technological innovations.

Veblen's focus on the conflict between the instinct of workmanship and predatory and pecuniary instincts is often referred to as the instrumental-ceremonial dichotomy. Ayres (1938) in particular reinforced the tendency of the OIE to focus on the past binding and ceremonial aspects of institutions and on the scientific and progressive nature of instrumental reasoning. This dichotomy was, at one point in time, a core proposition of the OIE.

Most contemporary OIEs, however, recognize and accept that at least some institutions can promote and facilitate progressive change and that technology itself is an institution. This rethinking of the ceremonial-instrumental dichotomy is also reflected in the incorporation of Karl Polanyi's (1944) dichotomy between habitation and improvement. Polanyi noted that the need for social protection may actually serve a noninvidious purpose. Some improvements destroy livelihoods and reinforce invidious distinctions while others promote the life process. So the distinction might better be thought of in terms of "invidious versus noninvidious."

One OIE who had a more positive understanding of the role of institutions is J. R. Commons (Commons, 1970; Wunder & Kemp, 2008). Commons in particular focused on the need for order in society and thus addressed the evolution of legal systems and the state. Commons's theory is primarily microevolutionary insofar as he focuses on the evolution of legal arrangements and shifting power alignments in modern industrial states. Commons is not as critical of existing arrangements as Veblen. Institutions,

including the state, in Commons's view, are clearly both necessary and potentially beneficial. For example, with the rise of big business, labor conflict, and the problems inherent in the business cycle, there is a need for a strong state to manage this conflict. At the same time, Commons developed a theory of the business cycle that has strong elements in common with some of Keynes's analysis.

The Veblenian strand as expressed by Commons is, by the standards of American politics, moderately left of center in that it expresses support for much of the regulatory framework and expanded role of government in managing the business cycle that came out of the New Deal and the publication of Keynes's (1936) *The General Theory of Employment, Interest and Money.* Not surprisingly, a number of OIE economists have begun to attempt to synthesize OIE and Keynes, relying to a large degree on the work of Hyman Minsky (1982). This project, often referred to as PKI (post-Keynesian institutionalism), is microevolutionary in nature in that it focuses on the problems of financial instability created by financial innovation and deregulation. The goal of PKI is wisely managed capitalism (Whalen, 2008). PKI clearly has a focus on the possibility of designing effective institutions, which logically implies that at least some institutions can embody instrumental reasoning.

In contrast to the direction taken by some OIEs, Hodgson (2004) has argued that Veblen's focus on technological thinking and the Commons-Ayres trend in OIE was a wrong turn for OIE. He has sought to revivify OIE by reinterpreting Veblenian economics as generalized Darwinism. Generalized Darwinism, according to Hodgson, generalizes the basic principles of Darwin's biological theory of evolution (inheritance, variation, and selection) to sociocultural evolution. In Hodgson's view, the mechanisms of inheritance, variation, and selection are not just analogies or metaphors to explain outcomes in social evolution—they are ontological principles that describe any entity that evolves. As noted above, because institutions and organizational routines are inherited through cultural learning and vary, they are subject to selection. Social evolution is therefore a special case of the more general case of evolution.

However, Hodgson (2004) also acknowledges that human agents are purposive and that culture is an emergent phenomenon. So Hodgson is not seeking to biologize social inequality or to reduce the social sciences to genetic principles such as inclusive fitness. Indeed, as Hodgson states, "more is needed" than just the principles of inheritance, variation, and natural selection. This would appear to be an understanding of how social institutions, in concert with instincts and human agency, generate outcomes in a complex, emergent process of social evolution. To this end, Hodgson has incorporated some elements of structure agency theory into his analysis.

Hodgson's program could be taken as an injunction to OIEs to build models of change that incorporate both Darwinian principles as well as more complex concepts of structure and agency. Hodgson has used this model to explain how changes in firm organization can be selected for or against by changes in market structure. So there are strong parallels between the work of Hodgson and that of Nelson and Winter (1982), who could notably be placed in either the OIE or NIE camp. As noted in the preceding section, Hodgson's view of evolutionary economics as "generalized Darwinism" is controversial, even among his fellow OIEs.

One competing strand of Veblenian economics is the radical strand as advocated by Bill Dugger (Dugger & Sherman, 2000). Dugger focuses on the role of technology, instrumental reasoning, and institutions as providing the capacity for improving the material condition of humans. The full application of instrumental reasoning, however, in Dugger's view is blocked by the key institutions of capitalism. These institutions are reinforced by ceremonial myths. Dugger also puts more emphasis on the social and ideological implications of the respective traditions and has been sharply critical of the NIE. He has also notably been instrumental in promoting dialogue between Marxists and OIEs and has often copublished works on sociocultural evolution with Howard Sherman. Dugger also tends to emphasize the non-Darwinian nature of sociocultural evolution.

## The New Institutionalists

It can be fairly argued that Adam Smith was the first evolutionary economist, even though his contributions predate any significant consideration of biological evolution by naturalists. Adam Smith provides an account of how an increasingly complex society arises out of the natural propensity of humans to truck, barter, and exchange (Fusfeld, 1977; Smith, 1776/1937). Ironically, some of Smith's concerns with specialization and division of labor, as well as the writings of another political economist, Thomas Malthus, influenced Darwin. Many Social Darwinists in the late nineteenth century drew on Darwinian reasoning to explain how competitive markets work and to justify social inequality. Some twentieth-century theorists such as Frederick Hayek and Larry Arnhart have tended to view the market as a natural outgrowth of human genetic endowments.

Taken as a whole, however, evolutionary explanations fell out of favor among economists in the twentieth century. In the late nineteenth century, the social sciences became increasingly fragmented, and the new field of economics increasingly lost its evolutionary focus. With the triumph of the standard competitive model in the mid-twentieth century, economics became narrowly focused on providing formal mathematical proofs of narrowly defined "how" questions. However, there are some signs that the standard competitive model is in the process of being displaced by game theory. There is also widespread recognition that it is necessary to supplement the standard competitive model with an evolutionary account. These developments have led to an increased acceptance of evolutionary explanations among mainstream economists and renewed attention to the importance of institutions in framing economic outcomes.

Some strands of the NIE, particularly the version espoused by Coase (1974) and Williamson (1985), view institutions primarily as providing "solutions" to the problems of asymmetric information and transactions costs. This strand of NIE does not significantly challenge the standard competitive model or its underlying behavioral assumptions. To the contrary, it is a complement to the standard competitive model. It is also to a large degree a micro-oriented theory of sociocultural evolution.

A more dynamic view of economic evolution is that of Joseph Schumpeter (1908, 1950). Schumpeter focused on the individual entrepreneur and his role in promoting technological innovation. This technological innovation disturbs the equilibrium and leads to gales of creative destruction. However, with the rise of the modern, bureaucratically organized firm, the role of the entrepreneur was lessened, leading to a static and moribund organization. Schumpeter thought that this would eventually lead to the destruction of capitalism, an outcome that, in contrast to Marx, Schumpeter viewed in a negative way. Schumpeter, however, drew a strong distinction between statics, exemplified by the Walrasian model of his day, and dynamics, exemplified by theories of economic evolution. Thus, "dynamics" was intended to complement "statics" (Andersen, 2008). Many contemporary mainstream models of economic growth, often referred to as new growth theory, explicitly incorporate Schumpeterian analysis.

Some of the richness of Schumpeter's focus on technological innovation as gales of creative destruction has been recaptured by the economic historian Joel Mokyr (1990) in his masterful work on technological progress. Mokyr adapts Gould's concept of "punctuated equilibrium" to the history of technology. He also draws a distinction between invention (the rise of new techniques and processes) and innovation (the spread of these new techniques). The Industrial Revolution, in Mokyr's view, is ongoing but is nevertheless a clear instance of a dramatic change in technological and social organization. Similarly, the work of Nelson and Winter (1982), previously cited, which acknowledges the contributions of Veblen, can also be considered neo-Schumpeterian. There are, it should be noted, significant parallels between Marx, Schumpeter, and Veblen, as well as differences.

The most prominent and most successful NIE, of course, is Douglas North. North's career has spanned several decades, during which his contributions to multiple fields in economics have been voluminous. Notably, North's own views themselves have undergone significant evolution. North's (1981) earlier work on economic evolution was an application of the work of Coase (1974) and Williamson (1985) to the problem of economic evolution and did not significantly challenge the standard competitive model. North viewed economic evolution as taking place due to changing resource constraints in response to the growth in population as rational agents calculated the marginal costs and marginal benefits of shifting from foraging to farming.

North's later work (1990, 1991, 1994), however, has challenged many aspects of the standard competitive model. North has focused specifically on the role institutions play in cognitive framing of decision making. Notably, North has explicitly abandoned the theory of strong rational choice in favor of models of human behavior that focus on the limited ability of humans to obtain, process, and act on information. In most textbook models of market behavior, price is the primary means of providing information. But in North's view of markets, information encompasses much more than price. In addition, norms, values, and ideology can blunt the ability of humans to obtain and interpret some information. North is not arguing that humans are "irrational" as his approach still logically implies some degree of calculation and conscious decision making based on self-interest. But he has abandoned the strong view of rationality, which implies humans are lightning rods of hedonic calculation. In that sense, his view of human behavior is much closer to that of the Austrians in focusing on the purposiveness of human behavior.

For the most part, North tends to see institutions as constraints on human action, though he acknowledges that institutions can provide incentives both in terms of the things we actually do, as well as the things that we do not do. Thus, institutions that reward innovative behavior, risk seeking, and trade will lead to efficient outcomes. Institutions that reward rent seeking and prohibit innovation and trade will lead to inefficient outcomes. Once an institutional structure is set, there is a strong degree of inertia that perpetuates the existing institutional structure. In other words, evolutionary paths, in North's view, tend to be path dependent. Clearly, the kinds of institutions in North's view that promote efficient outcomes are those that clearly define the rules of the game in favor of the operation of markets. This does not necessarily imply laissez-faire as the state may still be necessary to perform multiple functions. It does serve to distinguish between states, such as Great Britain in the seventeenth and eighteenth centuries or South Korea in the past several decades, that were able to define an institutional framework that promoted innovation and growth as opposed to states such as Spain in the sixteenth and seventeenth centuries or in the Congo (Zaire) today that destroy any incentive for innovation and economic growth.

This raises two very interesting questions. How does a particular type of path become established, and how does it change? North's explanation is one that is rooted in a metaphor of variation and selection. Greater variation will allow for a higher probability that a particular path will be successful. Greater centralization will reduce variation and increase the chances that the state will adopt or promote institutions that blunt technological and social innovation. North explains the greater success of Europe versus the rest of the world as a result of the relative decentralization of Europe in the early modern period. Arbitrary authoritarian states that destroyed incentives for growth such as Spain existed. But Spain was unable to impose its will on Europe

or on the emerging world market. Consequently, this enabled states such as England, where the power of the Crown became limited as Parliament enacted laws to protect commercial interests and innovation, to industrialize rapidly and emerge as world leaders. These contrasting paths were transferred to the New World. The United States inherited and successfully modified the institutional framework of Britain and therefore developed. Latin America inherited and failed to successfully modify the institutional framework of absolutist Spain and developed much more slowly.

## Whither Evolutionary Economics?

Evolutionary economics clearly has a future. Economists in general are becoming more attuned to the importance of understanding how humans organize the economy through institutions and how institutions change over time. This entails extensive borrowing of concepts from evolutionary biology and a reconsideration of the underlying behavioral assumptions of mainstream economics. Understanding how institutions permit or inhibit changes in technology, as well as how changes in technology in turn require changes in institutions, is a concern of all three schools of evolutionary economics. As NIE economists push the boundaries of the mainstream, at least some have increasingly asked heterodox questions, and a few have been willing to acknowledge heterodox contributions. Some Marxist and OIE scholars have also begun to note that at least some versions of NIE, if not necessarily entirely new, are at least genuinely institutional and evolutionary. Any grand synthesis seems distant, but there is at least a basis for further argumentation and even dialogue.

## References and Further Readings


Andersen, E. S. (2008). Appraising Schumpeter's "essence" after 100 years: From Walrasian economics to evolutionary economics. In K. K. Puranam & R. Kumar Jain B. (Eds.), *Evolutionary economics.* Hyderabad, India: ICFAI University Press.

Ayres, C. (1938). *The problem of economic order.* New York: Farrar and Rinehart.

Coase, R. H. (1974). The new institutional economics. *Journal of Institutional and Theoretical Economics, 140,* 229–231.

Commons, J. R. (1970). *The economics of collective action.* Madison: University of Wisconsin Press.

Cypher, J., & Dietz, J. (2004). *The process of economic development.* London: Routledge.

Dugger, B. (1988). Radical institutionalism: Basic concepts. *Review of Radical Political Economics, 20,* 1–20.

Dugger, B. (1995). Veblenian institutionalism. *Journal of Economic Issues, 29,* 1013–1027.

Dugger, B., & Sherman, H. (Eds.). (2000). *Reclaiming evolution: A dialogue between Marxism and institutionalism on social change.* London: Routledge.

Dugger, B., & Sherman, H. (Eds.). (2003). *Evolutionary theory in the social sciences: Vol. 1. Early foundations and later contributions.* London: Routledge.

Fusfeld, D. R. (1977). The development of economic institutions. *Journal of Economic Issues, 11,* 743–784.

Harris, M. (1997). *Culture, people, nature* (7th ed.). New York: Addison Wesley Longman.

Hodgson, G. M. (2004). *The evolution of institutional economics: Agency, structure and Darwinism in American institutionalism.* London: Routledge.

Hodgson, G. M. (Ed.). (2007a). *The evolution of economic institutions: A critical reader.* Cheltenham, UK: Edward Elgar.

Hodgson, G. M. (2007b). Introduction. In G. M. Hodgson (Ed.), *The evolution of economic institutions: A critical reader* (pp. 1–15). Cheltenham, UK: Edward Elgar.

Hodgson, G. M., & Knudsen, T. (2006). Why we need a generalized Darwinism and why a generalized Darwinism is not enough. *Journal of Economic Behavior and Organization, 61,* 1–19.

Keynes, J. M. (1936). *The general theory of employment, interest and money.* Cambridge, UK: Cambridge University Press.

Mayr, E. (2001). *What evolution is.* New York: Basic Books.

Mayr, E. (2004). *What makes biology unique?* Cambridge, UK: Cambridge University Press.

Minsky, H. (1982). *Can "it" happen again? Essays on instability and finance.* Armonk, NY: M. E. Sharpe.

Mokyr, J. (1990). *The lever of riches: Technological creativity and economic progress.* New York: Oxford University Press.

Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change.* Cambridge, MA: Harvard University Press.

North, D. (1981). *Structure and change in economic history.* New York: W. W. Norton.

North, D. (1990). *Institutions, institutional change and economic performance.* Cambridge, UK: Cambridge University Press.

North, D. (1991). Institutions. *Journal of Economic Perspectives, 5,* 97–112.

North, D. (1994). Economic performance through time. *American Economic Review, 84,* 359–367.

Poirot, C. S. (2007). How can institutional economics be an evolutionary science? *Journal of Economic Issues, 51,* 155–179.

Polanyi, K. (1944). *The great transformation.* New York: Farrar and Rinehart.

Schumpeter, J. (1908). *Das Wesen und der Haupterhault der theoretischen Nationolokonomie* [The nature and essence of theoretical economics]. Leipzig, Germany: Duncker und Humboldt.

Schumpeter, J. (1950). *Capitalism, socialism and democracy* (3rd ed.). New York: Harper.

Sherman, H. (1995). *Reinventing Marxism.* Baltimore: Johns Hopkins University Press.

Sherman, H. (2006). *How society makes itself.* New York: M. E. Sharpe.

Smith, A. (1937). *An enquiry into the nature and causes of the wealth of nations.* New York: Modern Library. (Original work published 1776)

Veblen, T. (1898). Why is economics not an evolutionary science? *Quarterly Journal of Economics, 12,* 373–397.

Veblen, T. (1899). The preconceptions of economic science. *Quarterly Journal of Economics, 13,* 121–150.

Whalen, C. J. (2008). Toward wisely managed capitalism: Post-Keynesianism and the creative state. *Forum for Social Economics, 37,* 43–60.

Williamson, O. E. (1985). *The economic institutions of capitalism.* New York: Free Press.

Wolf, E. (1982). *Europe and the people without history.* Berkeley: University of California Press.

Wunder, T., & Kemp, T. (2008). Institutionalism and the state: Founding fathers re-examined. *Forum for Social Economics, 37,* 27–42.

# 91

# Matching Markets

Thomas Gall
*University of Bonn*

A matching market assigns objects to individuals, or individuals to each other. Typically, the different objects are indivisible, and individuals differ in how much they value each of them, so that the assignment has important implications for the well-being of the individuals. Moreover, relevant applications involve markets where the use of monetary payments is limited or infeasible, such as public school choice, assignment of graduate students, or the exchange of live-donor kidneys for transplantation. In these markets, exhausting all opportunities for mutually beneficial exchange with the limited means available is important for the well-being and, in the case of the last example, the health of those involved. This chapter will demonstrate how economic theory can offer some guidance for the design of markets in order to solve such problems of assignment.

Several problems may arise in assignment problems that impede the attainment of a satisfactory outcome, where "satisfactory" could refer to Pareto efficiency or to other welfare criteria. Best understood among these problems are unraveling, strategic behavior and a failure to arrive at a stable allocation; they will be defined and discussed later in greater detail. Indeed, a growing body of economic research on market design is concerned with developing mechanisms that ensure that outcomes with desirable welfare properties are reached, while ensuring that individuals have adequate incentives to participate and to truthfully reveal their preferences over how much they value the objects to be assigned. Two such mechanisms, the Gale-Shapley mechanism and the top trading cycles mechanism, will be presented here, as well as applications to assigning students to colleges and schools and to the exchange of live-donor kidneys for transplantation.

This chapter is organized as follows. The next section gives an introduction to the theory of matching markets, discussing characteristics that typically distinguish matching from competitive markets, such as heterogeneity, indivisibility, and a lack of market prices. Then the concepts of stability, strategy-proofness, and optimality of an outcome are presented. Also, two algorithms that can be used to achieve outcomes with these properties are briefly discussed. To illustrate the practical relevance of matching markets, three applications of real-world assignment problems and the methods that have been employed to solve them are then described. A short discussion of policy considerations and a brief survey of further topics in matching market theory follow, and several areas for future research are reviewed that promise to eventually generate interesting and much-needed results. The chapter concludes with a summary. Also included is a detailed list of references for the interested reader.

## Theory

The following gives a brief introduction in the theory of matching markets, with an emphasis on stable outcomes and methods to implement such outcomes, while ensuring that participants have no interest to misrepresent their preferences. This is demonstrated both in a marriage market model and a housing market setup.

### Principles and Terminology

Assigning individuals to objects or, to a lesser extent, to other individuals appears to be a feature common to most markets. There are, however, some characteristics

that typically differentiate matching markets from competitive markets in general. For one thing, objects and individuals in a matching market are usually heterogeneous and indivisible. This means that "goods" on a matching market (e.g., individuals) are typically supplied and demanded in quantities of 1 that cannot be further broken down. In a competitive market, a good that is demanded by many traders will be divided to satisfy the traders' demands at the market price.

Indeed, the use of the word *market* seems to imply that a market price will be used to equate demand and supply for goods or individuals. While this might often be a viable method of assignment (e.g., one used in auctions), there are a number of applications where reaching a market price might be infeasible. This is true, for example, when prices are regulated, as in the choice of public schools or of dormitories at a university where places are provided essentially for free and as in admission into certain professions with rigid wage-setting conventions. This is also true when prices and monetary payments cannot be used on ethical grounds, such as for exchanges of kidneys from live donors. Moreover, a market price may not be informative about preferences, for instance, when some participants in a matching market are subject to credit constraints, as in the cases of university choice and the job market. In an extreme case, when individuals have no access to loans but need to make a sizable investment in tuition fees or special training, personal wealth will at least partially determine individual rankings of schools or jobs. Absence or limitations of a market price and monetary payments is often referred to as a situation with *nontransferable utility.*

If a matching market is not cleared by the market price, its outcome, also referred to as the matching allocation, will depend crucially on the method used to assign individuals to objects. For instance, the order in which individuals are allowed to choose may matter, as earlier choosers have a better chance that their preferred match is still on the market (think of the drafts in American professional sports leagues). Hence, a relevant issue is how to evaluate and compare different conceivable outcomes of matching markets. One important concern is that a matching allocation should be final in the sense that all feasible profitable exchanges are exhausted and there is no mutually profitable opportunity for rematching among individuals. This property is called stability. If an assignment scheme violates stability, individuals who expect that they will have such a profitable rematching opportunity given an assignment might decide they are better off bypassing the assignment scheme, for instance, by attempting to secure a favorable match before the market takes place. This is known as unraveling of a matching market and can lead to quite unsatisfactory matching allocations (see, e.g., Li & Rosen, 1998; Roth & Xing, 1994).

Having identified a desirable property of allocations such as stability, the next step is to find an assignment scheme that actually reaches a stable outcome. Such an assignment scheme is often simply an algorithm specifying step-by-step instructions for the assignment based on the preferences of market participants. Because the algorithm is based on information announced by participants (i.e., their stated preferences), quality of outcomes can be evaluated only with respect to reported information. Therefore, it is of great importance to elicit truthful revelation of information by individuals. Devising algorithms, or mechanisms, that reach desirable outcomes from a social point of view while ensuring that agents do not benefit by strategically misrepresenting their preferences is the object of the field of market design (see Roth, 2002, for an overview).

Matching market mechanisms are usually centralized in the sense that assignments are generated by a central matchmaker or a clearinghouse. This is in contrast to what is typically the case in the literature on search and matching in the labor market. There matching is decentralized, often meaning that participants in the market meet and match randomly.

## Two-Sided Matching Markets

The simplest instance of a matching market is the *marriage market* model. In a marriage market, economic agents belong to one of two different groups, called the *market sides* (e.g., men and women or students and colleges). Suppose, following the first example, that there are $n$ women to be assigned to $m$ men, such that every woman is matched to one man or stays solitary, and likewise every man is matched to one woman or stays solitary. That is, there are two disjoint market sides, and an individual on one market side can only be assigned to a member of the opposite market side. The restriction of allowable matches to those between two market sides is the defining characteristic of a *two-sided* matching market, as opposed to a *one-sided* matching market where there is only one market side in the sense that a priori any individual is free to match with any other individual. Furthermore, here each man is matched to at most one woman, and vice versa, so that matching is *one to one.* In the context of students and colleges, typically many students may be assigned to one college, so that matching is *many to one.*

Men and women have preferences in the form of complete rankings of all individuals from the other market side that they would prefer to be matched with over staying solitary. Suppose that preferences are strict, so that no individual is exactly indifferent between being matched with any two members of the opposite side (or between being matched with any member of the opposite side and staying solitary).

## Stability of a Matching Allocation

As was mentioned above, a desirable property for an outcome of a matching market is that it is final in the sense that all mutually profitable matches have been exhausted.

To capture this, say that an outcome can be *blocked* if there is an individual or a pair of individuals who can make themselves strictly better off by altering their own assignment when mutual consent is required. More precisely, an outcome can be blocked by an individual, if he or she is assigned to somebody in the outcome but would prefer to stay solitary. An outcome can be blocked by a pair, if a pair of individuals, one from each side, is not assigned to each other in the outcome but would both strictly prefer this to their actual assignment in the outcome. Consequently, an outcome is called *stable* if there is neither an individual nor a pair that can block the outcome. That is, to borrow the image of a marriage market once more, given a stable outcome, no mutually desired marriage is left un-brokered, and nobody will demand a divorce. This illustrates well the appeal of a stable allocation: In a stable allocation, an individual cannot gain by choosing differently because all matches preferred to the current one would not accept a proposal. For this reason, stability is also sometimes said to eliminate justified envy, where envy is considered only justified when an envied match would actually accept a proposal.

## The Gale-Shapley Algorithm

Fortunately, existence of a stable outcome in a marriage market is always guaranteed, as David Gale and Lloyd Shapley (1962) found. Since their proof is highly instructive for the understanding of a great variety of assignment problems, a sketch of it is given here. They prove existence constructively by proposing a very simple algorithm that never fails to achieve a stable matching allocation in a marriage market. A matching algorithm provides a detailed set of instructions for how to assign and possibly rematch individuals on both market sides. It proceeds sequentially in steps until a stopping condition is met, for instance, when there is no activity in the current step. Then it terminates, and the current assignment becomes the matching allocation. The *Gale-Shapley algorithm* is as follows:

*Step 1*

(i) Each man proposes to his first choice, if this is preferred to remaining solitary. Otherwise, he does not propose.

(ii) Each woman tentatively accepts the most preferred proposal made to her, or stays solitary if that is preferred, and rejects all other proposals.

*Step k*

(i) Each man who has been rejected at step $k - 1$ proposes to his $k$th choice, if this is preferred to remaining solitary. Otherwise, he does not propose.

(ii) Each woman tentatively accepts the most preferred proposal made to her (out of all proposals in the current step and her most recent previous tentative acceptance), or stays solitary if that is preferred, and rejects all other proposals (including the supplanted previous tentative acceptance if applicable).

*End*

The algorithm stops when a step is reached in which no proposals are made. All tentative acceptances by women become permanent, and the resulting matches constitute the outcome.

The key characteristic of this algorithm is that women hold on to the best proposals they receive but are free to reject any previous proposal should a better one arrive later on. This means acceptances by the women are not binding as long the algorithm has not terminated. This is why this algorithm is also sometimes called the *deferred acceptance* algorithm.

To verify that a marriage market has a stable outcome, one can proceed in two steps: It has to be confirmed first that the algorithm always produces an outcome and second that the outcome produced is indeed stable. For the first statement, one simply notes that both men and women are finite in number. Therefore, after a finite time, the algorithm must have exhausted all entries in all individual rankings so that nobody proposes anymore, or, if not, there must have been a step when nobody proposed. Moreover, the outcome of the algorithm does not depend on the order in which men are allowed to propose in each step or on the order in which women are allowed to decide which proposals to reject. It does, however, depend on whether men or women propose, as will become clear below.

That the matching outcome produced by a Gale-Shapley algorithm must be stable can be seen as follows. Suppose there are a man and a woman who can block the outcome. Then, under the rules of the algorithm, the man must necessarily have proposed to the woman before and been rejected. But then, again according to the rules of the algorithm, the woman cannot prefer a match with this man to the match she is assigned in the outcome. Hence, the initial assumption that there is a pair that can block the outcome must be false. Neither can any individual block the outcome since men only propose to women if this is preferred to staying solitary, and women only hold on to such proposals that are preferred to staying solitary. Therefore, neither an individual nor a pair can block the outcome, and so the outcome of a Gale-Shapley algorithm is stable.

## Optimal Matching Allocations

There may, however, be more stable outcomes than just the one arrived at by this algorithm. Furthermore, different stable outcomes may differ widely in how well-off the individuals find themselves. Gale and Shapley (1962) provide a useful result: There is always a stable outcome that is at least weakly preferred by all men to all other stable outcomes, which is typically called the *men-optimal* (or student-optimal in the case of college admission) stable outcome. Likewise, there is always a stable outcome that all women at least weakly prefer to all other stable outcomes, which is called the *women-optimal* stable outcome. Moreover, Gale and Shapley show that the men-optimal

outcome is reached by a deferred acceptance algorithm in which the men propose (as outlined above) while, if instead the roles are reversed and women propose, such an algorithm reaches the women-optimal outcome.

This result has an immediate implication for how policy makers would want to implement the algorithm if one market side were of more concern than the other. For instance, suppose one market side consists of organizations rather than individuals—say, in a student-college interpretation of the marriage market, where men can be relabeled as students and women as colleges. Then the student-optimal outcome seems clearly desirable from a social point of view, when assuming that the well-being of colleges is not a crucial social concern. This suggests that it would be desirable from a social point of view to implement the algorithm such that it is the students who make the proposals.

## Truthful Information Revelation

Using the Gale-Shapley algorithm, a social planner can implement a stable outcome and furthermore, simply by choosing which market side will play what role, can implement the stable outcome that is optimal for a given market side. However, the social planner must apply the algorithm on the basis of submitted preference rankings from both market sides. This means that the outcome that is reached will be stable only with respect to the stated preferences. If individuals do not report their preferences truthfully, however, the outcome might not be stable with respect to the true preferences, which may lead to Pareto inefficiency, secondary markets, or unraveling. Therefore, a further desirable property of any assignment mechanism is to ensure that individual participants cannot gain by misrepresenting their preferences. An assignment mechanism is called *strategy-proof* if it ensures that, for all possible combinations of individual strict preferences, participants at least weakly prefer to reveal information on their preferences truthfully. The interested reader may consult the textbook on two-sided matching by Roth and Sotomayor (1990) for a more exhaustive treatment of the following results and many others.

First, a negative result has to be mentioned: There is no assignment mechanism that generally implements a stable outcome in a marriage market and is strategy-proof (Roth, 1982). To see this, suppose men- and women-optimal outcomes do not coincide, and an algorithm is in place where men propose (as described above). Then all women prefer an outcome other than is reached by the algorithm. In particular, the following situation may arise. A woman misrepresents her preferences by eliminating her match in the men-optimal outcome from her submitted list. Then this man is rejected and tentatively assigned to another woman who prefers that man to her match in the men-optimal outcome. The newly rejected man in turn proposes to the first woman (who lied) and is tentatively accepted, if this woman prefers him to her match in the men-optimal outcome. That is, women exchange men such that all participating women are better off and all exchanged men worse off. Because this possibility cannot be precluded in general, strategy-proofness cannot be ensured for both market sides.

Lester Dubins and David Freedman (1981) present a slightly more encouraging result in finding that a mechanism that yields the men-optimal outcome on a marriage market ensures that no man can gain by misrepresenting his true preferences. That is, such a mechanism is strategy-proof among the men. This must be the case because by misrepresenting preferences, a man may induce a different stable outcome, but the men-optimal outcome is preferred by all men among all stable outcomes. A similar property holds for a mechanism inducing a women-optimal outcome: It is strategy-proof among the women. This is in fact a very useful result in case one is interested in a matching market where participants on one market side can be considered to be not self-interested. This may be the case, for instance, when assigning pupils to schools, students to universities or courses, or airplanes to landing slots. Some caution is appropriate, however, when the matching is many to one, as the result on strategy-proofness for one market side holds only for the market side having one partner each. To illustrate this, consider a school choice problem where schools have multiple places. A matching mechanism yielding the pupil-optimal allocation is strategy-proof among the pupils, but a matching mechanism yielding the school-optimal allocation is not necessarily strategy-proof among the schools.

## One-Sided Markets

In contrast to the two-sided marriage market discussed so far, an individual participating in a one-sided matching market may be assigned to any other individual participating in that market. Examples are team or group formation or marketplaces where indivisible objects are exchanged, such as markets for houses or dorm rooms. In these markets, every individual may initiate a trade with any other individual or group of individuals. The basic *housing market* model due to Lloyd Shapley and Herbert Scarf (1974) considers an exchange economy where $n$ individuals trade in an indivisible good, say houses. Each individual owns one house when entering the market, has need of exactly one house, and possesses strict preferences over all existing houses. Potential matches in this market are between individuals and indicate house trades.

An outcome in this market is an allocation of houses among individuals such that each individual holds at most one house. That is, potential trades, or matches, among (groups of) individuals generate an allocation of houses to individuals, the market outcome. Note that also a school choice matching market could potentially be formulated as a one-sided market, if pupils are endowed with a place at certain school. Shapley and Scarf (1974) use the *core* as a

solution concept to determine the outcome. An outcome is in the core if there is no group of individuals (of size one or greater) that could make every group member weakly and at least one strictly better off by reallocating the houses owned by group members among members of the group. This is equivalent to saying that an outcome exhausts all opportunities for mutually beneficial trade among any number of individuals (not only between pairs of members of opposing market sides).

### The Top Trading Cycles Algorithm

When all individuals have strict preferences, there is always at least one matching allocation in the core (Roth & Postlewaite, 1977). Shapley and Scarf (1974) show that an outcome in the core can be reached by following a specific algorithm, the *top trading cycles* algorithm, which they attribute to David Gale. Before presenting the algorithm, it will be useful to explain what is meant by *cycles*. A cycle is a *sequence of individuals* such that each individual is followed by, or points to, the owner of his or her most preferred house, and the last individual in the sequence most prefers the first individual's house. A cycle may consist of one individual only. Note that, among a finite set of individuals who have strict preferences, there must always be at least one cycle. For cycles of more than one individual, if all members of a cycle give their houses to their successors in the sequence and the last individual to the first, all individuals in the cycle will be strictly better off. That is, cycles identify opportunities for mutually strictly beneficial trades in groups. The top trading cycles algorithm exhausts all such opportunities and works as follows.

*Step 1*

Let all individuals point to the owners of their most preferred houses. Identify all cycles, effect the implied exchange of houses, and remove the individuals in the cycles from the market.

*Step k*

Let all individuals that remain in the market after step $k-1$ point to the owners of their most preferred houses. Identify all cycles, effect the implied exchange of houses, and remove the individuals in the cycles from the market.

*End*

The algorithm stops when there are no more individuals in the market.

This means that the top trading cycles algorithm iteratively identifies all trading cycles, executes the trades, and removes the individuals in the cycles until the market is cleared. The outcome generated by this algorithm is in the core for all housing markets (Shapley & Scarf, 1974). That is, given a matching allocation resulting from this mechanism, there is no opportunity for mutually strictly beneficial exchange among any group of agents. This means the outcome is Pareto efficient and essentially replicates the competitive equilibrium allocation (Roth & Postlewaite, 1977). Hence, from a social planner's perspective, outcomes of a top trading cycles mechanism on a housing market have highly desirable properties. Moreover, the mechanism does not require monetary payments and can therefore be deployed in situations in which competitive market prices may not be feasible. To make it a desirable mechanism, it should also ensure that participants have no incentive to misreport their preferences. Roth (1982) shows that this is indeed the case.

## Applications

In the following, a number of real-world assignment problems are presented, along with the matching markets that have arisen in correspondence. The aim of the following presentation is twofold. On one hand, assignment mechanisms that have been employed on matching markets are described and analyzed with respect to shortcomings such as unraveling or preference misrepresentation. On the other hand, possible remedies, or policies, are presented to amend any such shortcomings identified in the analysis.

### The Market for Medical Interns and Residents

The market for medical interns and residents in the United States is a two-sided matching market. On one side are colleges offering internship and residency positions, and on the other side are graduate students seeking such positions. Both colleges and students are very heterogeneous in academic quality, both are indivisible (as part-time employment is usually infeasible), and wages for students are fixed before the application process begins. Alvin Roth (1984) gives an account both of the problems encountered in this market and the theoretical reasons. Before a viable assignment mechanism was introduced in 1951, matching was conducted in a de-centralized manner by private initiatives of market participants. Because the number of positions exceeded that of students, colleges had an incentive to try to preempt competitors in securing good candidates by making offers earlier. Indeed, whereas students tended to obtain a position about half a year before graduation in the 1930s, they were typically offered a contract about 2 full years before graduation by 1944. Such early timing, of course, forgoes a lot of information on the quality of candidates, which is very likely to adversely influence the quality of the matches. This is an instance of unraveling. Others include markets for law clerks and gastroenterologists (see Avery, Jolls, Posner, & Roth, 2001; Niederle, Proctor, & Roth, 2006).

In 1945, an attempt was made to remedy the shortcomings by disclosing information on candidates only shortly before graduation. This prevented unraveling but created another problem, as offers and acceptances were binding.

It became costly for colleges to hold offers open for any amount of time because by the time an offer was rejected, the market was quite likely to be depleted. Not surprisingly, while offers remained open for 10 days in 1945, this time fell to about 10 minutes by 1950.

As this development was unsatisfactory, a clearinghouse was formed in 1951 that used a centralized mechanism, the National Internship Matching Program (NIMP). This mechanism works as follows. Students and colleges submit their preferences over the other market side in the form of rankings to the clearinghouse. An algorithm is then used to determine the matching outcome. The principal characteristic of the algorithm employed is that it uses deferred acceptance, as with the Gale-Shapley algorithm. Therefore, the outcome reached by the NIMP is stable with respect to the stated rankings by students and colleges (Roth, 1984). Hence, given the outcome, no participant can obtain a better match—that is, the allocation is envy free—and there should be no reason for a secondary market to emerge. This may explain why the program, although participation was voluntary, was quite successful. Indeed, the mechanism is still in use today, albeit renamed the National Residency Matching Program (NRMP) and in a slightly modified form. Some changes were designed to accommodate concerns such as differences in medical training programs and the desire of couples to end up in similar locations. More significantly, the algorithm was changed from a college-proposing to a student-proposing format to ensure student optimality of the matching outcomes after complaints arose that students had been treated badly. Potential gains to students from misrepresenting their preferences appear to be small in the NRMP, however (see Roth & Peranson, 1999, for further details).

## Public School Choice

The assignment of pupils to public schools provides another informative real-world example of a two-sided matching market. In many U.S. school districts, the assignment mechanism used to match pupils and schools often fails to implement a stable outcome, which can generate incentives to misreport preferences. For instance, the city of Boston employed the following mechanism to assign pupils to public schools in the years 1999–2005. Parents submitted a ranking of their top schools. Schools assigned priority to pupils who lived within walking distance, higher priority to those who had a sibling already enrolled, and highest priority to pupils who satisfied both criteria. Further details can be found in the studies by Abdulkadiroglu and Sönmez (2003) and Ergin and Sönmez (2006). The so-called Boston Student Assignment Mechanism (BSAM) proceeds as follows.

*Step 1*

For each school, pupils who have listed that school as their first choice are assigned in order of priority while the school still has free capacity. Ties are broken using a random procedure (such as flipping a coin).

*Step k*

For each school, pupils who have not been matched to a school in a previous step and who have listed that school as their *k*th choice are assigned in order of priority while the school still has free capacity. Ties are broken using a random procedure.

*End*

The algorithm stops, when either all pupils are assigned or there is no school left with remaining places.

The key feature of this algorithm is that assignments at each step are permanent—that is, when a school accepts a pupil, this acceptance is binding. Note that this is in contrast to deferred acceptance, which is characteristic of the Gale-Shapley algorithm. Ultimately, the fact that acceptances are permanent gives pupils an opportunity for strategic behavior. This implies in turn that the BSAM is not strategy-proof. This is because being rejected by one's top-ranked school may waste one's priority at another school if the other school is over-demanded and fills to capacity in a previous step. For example, if some pupil most prefers an over-demanded school at which he or she does not have priority but does have priority at his or her second most preferred school, it is strictly better for the pupil to list the second most preferred school in the first place on the submitted ranking to minimize the chances of being stuck with his or her third choice or worse. That is, under the BSAM, pupils can strictly gain by misrepresenting their preferences.

This observation has spawned a lively debate on possible implications for policy, in particular on whether to replace the existing mechanism by one that is strategy-proof, such as the Gale-Shapley algorithm. To offer theoretical guidance, Ergin and Sönmez (2006) investigate a game of school choice and find that the outcome of a pupil-optimal Gale-Shapley algorithm is more desirable from an efficiency perspective than the one of the BSAM (as long as the preferences of both market sides are strict). Furthermore, Adulkadiroglu, Pathak, Roth, and Sönmez (2006) and Pathak and Sönmez (2008) put forward the argument that a Gale-Shapley algorithm places lower cognitive and computational burden on participants than does the BSAM. This may be desirable, as evidence suggests that while some participants actually behave strategically, others announce their rankings sincerely. The BSAM systematically disadvantages these "naive" individuals if there are sophisticated players who behave strategically.

In 2006, public authorities changed the mechanism used in the assignment of pupils to public schools in Boston and introduced a version of the Gale-Shapley algorithm. New York City has also undergone a similar change in its assignment mechanism of pupils to high schools (see Abdulkadiroglu, Pathak, & Roth, 2005). There, a chief problem that had to be dealt with was strategic behavior of high schools.

### Kidney Exchange

As a final example, consider the problem of assigning kidneys from live donors to patients in need of transplantation. Often, patients who are in desperate need of a kidney transplant have a willing donor. (It is perfectly possible to live a healthy life with only one kidney.) Several compatibility conditions need to be met, however, to make a transplant feasible, such as having common blood types. Suppose that a patient has found a willing donor but is incompatible with that donor. Finding another patient who is compatible with the first donor and who has a live donor who is compatible with the first patient would make both patients better off (and presumably the donors, too). Obviously, it is of tremendous interest to identify all possible opportunities for mutually beneficial exchange because forgone opportunities to exchange will very likely result in the loss of human life.

Alvin Roth, Tayfun Sönmez, and Utku Ünver (2004) provide a theoretical analysis of this matching market and propose a mechanism for the exchange of kidneys from live donors. The key insight is that the market for kidney exchange is a one-sided matching market, similar to the housing market presented above. For this type of matching market, a mechanism is known that achieves a Pareto efficient allocation—that is, a matching outcome that exhausts all possible profitable trades—and is strategy-proof: the top trading cycles algorithm outlined above. Because not all patients necessarily have a donor to bring to the market, the appropriate theoretical model is a housing market with existing tenants, as studied by Abdulkadiroglu and Sönmez (1999), who explicitly allow for a waiting list. Roth et al. (2004) extend this theoretical work by allowing for the possibility of a waiting list for kidneys from deceased donors and derive a top trading cycles and chains (TTCC) mechanism that is Pareto efficient and strategy-proof.

The number of live-donor kidney exchanges that are actually carried out appears to be quite small relative to that which could be achieved by exhausting all potential for mutually beneficial exchange, pointing to a clear need for a centralized matching institution. In 2005, a clearinghouse (the New England Program for Kidney Exchange) for gathering data on donors and patients and generating the actual assignment was founded in New England (Roth, Sönmez, & Ünver, 2005, 2007). By 2008, this clearinghouse had already enabled a substantial number of two-way and even some three-way kidney exchanges. A similar program has since been started in Ohio.

## Further Issues

Recently, some further issues have been raised in public debate about some of the matching markets mentioned above. In the case of the market for medical residents, complaints have been made that the use of the NRMP depresses salaries for residents and fellows. An antitrust lawsuit in this matter was brought forward in 2002 but dismissed in 2004. Some theoretical support for the claim comes from Bulow and Levin (2006). They find that, when explicitly accounting for the fact that wage setting by colleges takes place before the matching market does, wages can indeed be depressed compared to the competitive outcome. Other studies (e.g., Kojima, 2007; Niederle, 2007) argue that this result hinges on particular assumptions that do not describe the real market for medical residents well.

As for school choice, recently it has been suggested that a version of the BSAM that modifies the random tie-breaking procedure would be able to better respect the intensity of pupils' preferences for different schools (see Abdulkadiroglu, Che, & Yasuda, 2008; Miralles, 2008). Thus, schools could be allocated to the pupils who have the highest valuations, which is different from merely ensuring that schools are assigned to pupils who prefer them to the next best school. This concern becomes more relevant when the pupils' preferences over schools are closely aligned, which seems to be the case in this market.

The cases mentioned above raise two additional issues. First, it is not clear whether the desirability of a mechanism is better judged based on whether it exhausts all potential mutually profitable exchanges or on whether it maximizes the sum of individuals' valuations of the outcome, that is, aggregate surplus. Aggregate surplus, also sometimes interpretable as output, may be a relevant criterion, for example, when analyzing economy-wide policies, such as labor market regulations and their effects on growth. Second, often there will be some means of compensation between partners, for instance, through the exchange of favors or gifts.

These issues have at least been partially addressed. When preferences are perfectly aligned—that is, the rankings of all individuals in the market agree—Becker (1974) shows that stable matching allocations need not maximize aggregate surplus when the possibility to compensate a partner in any form is completely excluded. Indeed, independently of which matching allocation maximizes aggregate surplus, a stable outcome in his model always assigns individuals whose attributes are more attractive to individuals or objects that have more attractive attributes. This is known as positive assortative matching. The general case, where compensation within a match is possible but costly in terms of surplus, has only recently been characterized. Legros and Newman (2002, 2007) provide conditions for assortative matching and confirm that stability and surplus maximization need not coincide.

When asymmetric information about individuals' attributes is a concern (e.g., concerning the productivity of firms and workers in labor markets), some form of segmentation and randomization of the matching within segments may be desirable (see, e.g., Jacquet & Tan, 2007; McAfee, 2002).

## Directions for Future Research

While the design of matching markets is already sufficiently well understood to offer useful contributions in many important assignment problems, there are a number of issues in which further theoretical progress would appear to be highly valuable. A first topic worth mentioning for being understudied is the forming of and properties of individuals' preferences. In large markets, constructing a complete preference ranking over every member of the other market side may be extremely costly because all potential matches would first need to be evaluated. In the labor market, for instance, this would require time-consuming interviews and assessments of and by all individual applicants and employers. A possible alternative may consist of designing a mechanism with a prematching stage in which participants indicate a crude preliminary ranking based on easily observable information (e.g., grades, public rankings). Such a mechanism has been tried out in the entry-level job market for economists since 2007. Whether this is indeed the optimal approach remains an open theoretical question.

Second, note that most of the theoretical results presented above rely on strict preferences, meaning that market participants cannot be indifferent between any two matches. This is indeed a limitation because the assumption cannot be easily discarded without affecting the theoretical properties of the mechanisms yet seems unreasonable in many relevant contexts. Some work has been done in identifying strategy-proof mechanisms for situations in which individuals may be indifferent between multiple matches. In such cases, the manner of breaking the tie and choosing the actual match may affect incentives for strategic behavior (see, e.g., Abdulkadiroglu, Pathak, & Roth, 2009; Erdil & Ergin, 2008; Miralles, 2008). Exploiting the possibility of indifference through the use of stochastic mechanisms—that is, matching mechanisms that use random procedures for assignment—appears to be a very promising field for future research. This is because such mechanisms allow respecting the intensity of individuals' preferences over matches (e.g., by setting odds to obtain a certain match that are acceptable only for individuals with intense preference for that match).

A third limitation of the theory of matching markets is that many of the theoretical results described above do not necessarily apply when individuals care not only about the match they obtain (e.g., the instructor, school, university, firm) but also about who else obtains the same match. This is, for instance, the case when thinking about couples searching jointly for jobs in the labor market. Although evidence suggests that a hands-on approach to market design works well in the market for medical residents (see Roth & Peranson, 1999), sound theoretical guidance for generating mechanisms that can also account for preferences over the outcomes of individuals on the same market side would be very welcome.

Another important area for future research is analyzing the effects of the anticipated outcome of a matching market (possibly generated by a mechanism) on individuals' behavior before entering the market. Often the attributes that the preferences of market participants are based on are subject to individual choices and investments. For example, the choice of how much education to acquire appears to affect individual outcomes in the labor market quite substantially. Some recent work suggests that, when partners in a match cannot compensate each other without incurring a loss in efficiency or when there is asymmetric information concerning attributes, stable outcomes need not maximize aggregate surplus and may distort individuals' prematching choices when these are based on anticipating the matching outcome (Gall, Legros, & Newman, 2006; Hoppe, Moldovanu, & Sela, 2009).

## Conclusion

There is a growing interest in and demand for applying economic theory to the design of mechanisms that solve real-world assignment problems. Cases in point are entry-level job markets, school choice, and exchange of live-donor kidneys. This chapter has presented an overview of such matching markets, typically characterized by heterogeneity and indivisibility of individuals and objects, as well as by limited possibilities for compensation of matching partners through side payments. A brief introduction to the theory of matching outlined important concepts such as stability of a matching allocation and strategy-proofness of an assignment mechanism. Two assignment mechanisms have been discussed in greater detail: the Gale-Shapley or deferred acceptance algorithm for two-sided matching markets and the top trading cycles algorithm for one-sided markets. Both algorithms satisfy two desired properties: stability of the resulting outcome and strategy-proofness.

The chapter proceeded to consider several real-world applications of matching markets in detail. In the market for medical residents and the choice of public schools, both two-sided matching markets, appropriate versions of the Gale-Shapley algorithm have been implemented with a high degree of success. An example of a one-sided matching market was provided in the exchange of kidneys for transplantation from live donors through a centralized facility using a version of the top trading cycles algorithm.

Further issues arise when it is in the social interest to choose a matching allocation that maximizes aggregate surplus. Because the stable outcome does not necessarily maximize aggregate surplus, a tension between surplus efficiency and stability may arise. In such cases, it is important to identify surplus-maximizing outcomes and to design and implement mechanisms that reliably reach such outcomes.

Finally, the chapter considered some relevant issues and questions that demand further investigation. The theory of matching markets needs to achieve higher levels of sophistication to provide results that are more applicable in

empirically relevant situations, such as when individuals are indifferent between multiple matches or have preferences concerning who else obtains the same match as they do. This is needed to be able to devise adequate matching mechanisms that can be successfully employed to address a wider range of assignment problems. Promising directions for future research lie in the use of stochastic mechanisms, which may use fine-tuned random procedures for the assignment, and in considering the dynamic environment of an assignment problem to evaluate effects on choices made prior to entering the matching market, particularly those that affect attributes relevant for the resulting matching allocation.

## References and Further Readings


Abdulkadiroglu, A., Che, Y., & Yasuda, Y. (2008). *Expanding "choice" in school choice* (Working paper). New York: Columbia University.

Abdulkadiroglu, A., Pathak, P., & Roth, A. E. (2005). The New York City high school match. *American Economic Review, 95,* 364–367.

Abdulkadiroglu, A., Pathak, P., & Roth, A. E. (2009). Strategy-proofness versus efficiency in matching with indifferences: Redesigning the NYC high school match. *American Economic Review, 99,* 1954–1978.

Abdulkadiroglu, A., Pathak, P., Roth, A. E., & Sönmez, T. (2006). *Changing the Boston school choice mechanism: Strategy-proofness as equal access* (NBER Working Paper 11965). Cambridge, MA: National Bureau of Economic Research.

Abdulkadiroglu, A., & Sönmez, T. (1999). House allocation with existing tenants. *Journal of Economic Theory, 88,* 233–260.

Abdulkadiroglu, A., & Sönmez, T. (2003). School choice: A mechanism design approach. *American Economic Review, 93,* 729–747.

Avery, C., Jolls, C., Posner, R. A., & Roth, A. E. (2001). The market for federal judicial law clerks. *University of Chicago Law Review, 68,* 793–902.

Becker, G. (1974). A theory of marriage: Part I. *Journal of Political Economy, 81,* 813–846.

Bulow, J., & Levin, J. (2006). Matching and price competition. *American Economic Review, 96,* 652–668.

Dubins, L. E., & Freedman, D. (1981). Machiavelli and the Gale-Shapley algorithm. *American Mathematical Monthly, 69*(3), 9–15.

Erdil, A., & Ergin, H. (2008). What's the matter with tie-breaking? Improving efficiency in school choice. *American Economic Review, 98,* 669–689.

Ergin, H., & Sönmez, T. (2006). Games of school choice under the Boston mechanism. *Journal of Public Economics, 90,* 215–237.

Gale, D., & Shapley, L. S. (1962). College admission and the stability of marriage. *American Mathematical Monthly, 69*(3), 9–15.

Gall, T., Legros, P., & Newman, A. F. (2006). The timing of education. *Journal of the European Economic Association, 4,* 427–435.

Hoppe, H., Moldovanu, B., & Sela, A. (2009). The theory of assortative matching based on costly signals. *Review of Economic Studies, 76,* 253–281.

Jacquet, N. L., & Tan, S. (2007). On the segmentation of markets. *Journal of Political Economy, 115,* 639–664.

Kojima, F. (2007). Matching and price competition: Comment. *American Economic Review, 97,* 1027–1031.

Legros, P., & Newman, A. (2002). Monotone matching in perfect and imperfect worlds. *Review of Economic Studies, 69,* 925–942.

Legros, P., & Newman, A. (2007). Beauty is a beast, frog is a prince: Assortative matching with nontransferabilities. *Econometrica, 75,* 1073–1102.

Li, H., & Rosen, S. (1998). Unraveling in matching markets. *American Economic Review, 88,* 878–889.

McAfee, P. (2002). Coarse matching. *Econometrica, 70,* 2025–2034.

Miralles, A. (2008). *School choice: The case for the Boston mechanism* (Working paper). Boston: Boston University.

Niederle, M. (2007). Competitive wages in a match with ordered contracts. *American Economic Review, 97,* 1957–1969.

Niederle, M., Proctor, D. D., & Roth, A. E. (2006). What will be needed for the new GI fellowship match to succeed? *Gastroenterology, 130,* 218–224.

Niederle, M., Roth, A. E., & Sönmez, T. (2008). Matching and market design. In S. N. Durlauf & L. E. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed., Vol. 5, pp. 436–445). New York: Palgrave Macmillan.

Pathak, P., & Sönmez, T. (2008). Leveling the playing field: Sincere and sophisticated players in the Boston mechanism. *American Economic Review, 98,* 1636–1652.

Roth, A. E. (1982). The economics of matching: Stability and incentives. *Mathematics of Operations Research, 7,* 617–628.

Roth, A. E. (1984). The evolution of the labor market for medical interns and residents: A case study in game theory. *Journal of Political Economy, 92,* 991–1016.

Roth, A. E. (2002). The economist as an engineer: Game theory, experimental economics and computation as tools of design economics. *Econometrica, 70,* 1341–1378.

Roth, A. E., & Peranson, E. (1999). The redesign of the matching market for American physicians. *American Economic Review, 89,* 748–780.

Roth, A. E., & Postlewaite, A. (1977). Weak versus strong domination in a market with indivisible goods. *Journal of Mathematical Economics, 4,* 131–137.

Roth, A. E., Sönmez, T., & Ünver, U. (2004). Kidney exchange. *Quarterly Journal of Economics, 119,* 457–488.

Roth, A. E., Sönmez, T., & Ünver, U. (2005). A kidney exchange clearinghouse in New England. *American Economic Review, 95,* 376–380.

Roth, A. E., Sönmez, T., & Ünver, U. (2007). Efficient kidney exchange: Coincidence of wants in markets with compatibility-based preferences. *American Economic Review, 97,* 828–851.

Roth, A. E., & Sotomayor, M. (1990). *Two-sided matching: A study in game-theoretic modeling and analysis.* Cambridge, UK: Cambridge University Press.

Roth, A. E., & Xing, X. (1994). Jumping the gun: Imperfections and institutions related to the timing of market transactions. *American Economic Review, 84,* 992–1044.

Shapley, L. S., & Scarf, H. (1974). On cores and indivisibility. *Journal of Mathematical Economy, 1,* 23–37.

# 92

# Beyond Make-or-Buy

## *Advances in Transaction Cost Economics*

Lyda S. Bigelow
*University of Utah*

*Thinking about problems of economic organization from a transaction cost economics point of view takes concerted effort. Once, however, a threshold of understanding is reached, the world of organization is "spontaneously" reordered. Applications abound. . . . The worldview of new students who get caught up in transaction cost reasoning is irreversibly altered.*

Oliver E. Williamson (1985, p. x)

Although Ronald Coase is often credited with inspiring a line of economic inquiry that came to be known as transaction cost economics, it is the work of Oliver Williamson, which appeared some 40 years later, that laid the foundation and provided the theoretical apparatus that has allowed scholars in this field to make headway on fundamental problems in law, economics, and the study of organizations. The magnitude of this contribution earned Williamson the Nobel Prize in Economics in 2010. This chapter sets out to familiarize the reader with central ideas and contributions of transaction cost economics and to illustrate the ongoing research trajectories that have emerged since Williamson set forth to expand on the implications of Coase's famous 1937 article. In addition, this chapter is written with the purpose of identifying recent areas of research that offer students the opportunity to consider possible empirical and theoretical extensions of transaction cost reasoning. Clearly, given the space constraints of a chapter, this survey cannot be exhaustive but will provide a basic outline of central insights and reflect the author's own assessment of the most promising new areas. The reader is asked to keep this caveat in mind in reading the chapter material.

The chapter is organized as follows: The first section provides a summary of the initial formulation of transaction cost economics with an emphasis on the above-mentioned "reordering" of economic organization; the next introduces the concept of discrete structural alignment and assesses the empirical evidence in predicting firm boundaries; the third section discusses contributions to problems beyond make-buy, focusing on strategic alliances, international business, and performance; the fourth section highlights emerging areas of research and research opportunities in the areas of contracts, technological evolution, and industry dynamics and concerns related to measurement and methodological issues; and the fifth section summarizes and concludes.

## Formulation of Transaction Cost Economics

Economics students are certainly aware of the Coase Theorem and the fact that Ronald Coase won the Nobel Prize for Economics in 1991. What they may not be aware of is the contribution Coase made to rethinking economic organization. In a seminal paper published in 1937, Coase proposed relaxing the assumption that transactions costs are zero, and once those costs are allowed to be positive, insights surrounding the efficacy of markets and the costs

of bureaucracy are altered. Coase is among a group of economists, who, beginning in the 1960s, sought to better understand institutions and the reasons for why we observe different forms of exchange. In essence, if markets work so well, why is there a need for firms at all? Why are contracts between firms specialized? With the publication of *Markets and Hierarchies,* Oliver Williamson (1975) began laying the theoretical foundation for modern transaction cost economics. This and his later work (e.g., Williamson, 1985, 1991, 1996), as well as the work of other economists (e.g., Grossman & Hart, 1986; Hart & Moore, 1990; Klein, Crawford, & Alchian, 1978), focused on addressing fundamental questions of the structure of economic exchange and how firms resolve the risks inherent in exchange in a world where there are no perfect governance solutions. (Note that throughout this chapter, the term *governance* is used to refer to the structure of economic exchange. This is how the term is used in the transaction cost literature. It is assumed that economic agents can choose and create governance structures that best protect them from the hazards of opportunistic behavior on the part of others, but no governance solution offers complete protection. More on this in the second section.)

Williamson (e.g., 1985, 1996) in particular espouses the view that there is broad applicability of this perspective. For example, he builds the case for using transaction cost economics to understand organizational, institutional, and nation-state exchange issues ranging from the familiar problems of contracting hazards to more unique phenomena such as the presence of company towns and the utilization of franchising. As an individual scholar, his appointment to three departments—economics, business, and law—while at the University of California, Berkeley, reflects his embodiment of this broad applicability of his work, based as it is on an interdisciplinary approach.

To best appreciate the insight/contribution of transaction cost economics, it is useful to put its development/appearance in historical context. Prior to the publication of *Markets and Hierarchies* in 1975, a dominant economic explanation for firms that used unusual strategies was to invoke a monopoly explanation. Indeed, Williamson (1985), in his introduction to *The Economic Institutions of Capitalism,* deploys the following quote from Coase (1972): "If an economist finds something—a business practice of one sort or another—that he does not understand, he looks for a monopoly explanation" (p. 67). Much of this focus was due to an era in which economists examined firm decisions with a high concern for antitrust and anticompetitive issues. Not surprisingly, this concern was driven, in part, by a period in which there were high levels of mergers and acquisitions. With the rise of the conglomerates in the 1960s, many economists, often those directly employed by the U.S. government antitrust divisions, sought to provide economic explanations for why these organization decisions led to anticompetitive practices revealed in unfair pricing, restraint of trade, and reduction in choice. While the concern was certainly warranted, Williamson sought to correct an overreliance on such antitrust explanations for firm behavior. His work was groundbreaking at the time because it carefully examined the same cases that had been deemed anticompetitive and offered an explanation that revealed transaction cost economizing and thus competitive rather than anticompetitive logic. For example, in his discussion of cable TV franchise bidding, Williamson (1985) elaborated a host of contracting problems that could have explained the behavior of the firm just as well as anticompetitive explanations.

Thus, Williamson illustrates that, indeed, there is an alternate logic underlying much of the actions that firms undertake. In other words, much of what we observe, within firms and within institutions that support economic activity in capitalist systems, may be explained with the understanding that the objective is to economize on transaction costs.

And what are transaction costs? These costs may be thought of as the economic equivalent of friction in physical systems. They are the costs of considering, crafting, negotiating, monitoring, and safeguarding contracts. Another powerful insight developed and elaborated by Williamson was to consider that most economic exchange can be undertaken either within or between firms (and later hybrids). Either way, understanding the ability of firms to economize on transaction costs requires recognizing that they must choose between alternate modes of organizing and that this choice must take into account both the features of the economic exchange, or transaction, as well as the features of the governance structure. In one of the most important statements in Williamson's *Economic Institutions of Capitalism,* he states, "The underlying viewpoint that informs the comparative study of issues of economic organization is this: Transaction costs are economized by assigning transactions (which differ in their attributes) to governance structures (the adaptive capacities and associated costs of which differ) in a discriminating way" (p. 18).

## Discrete Structural Alignment and Firm Boundaries

Beginning with his work in 1985 and perhaps best articulated in a paper that appeared in *Administrative Science Quarterly* in 1991, Williamson invoked the logic of discrete structural alignment. This is the theoretical apparatus that supports modern transaction cost theory. It begins with an understanding that there are different structural solutions to organizing economic exchange. These structural solutions, referred to as "governance," can be arrayed along a spectrum with markets on one end (arm's-length contracts, in which the identity of the parties to the exchange is not relevant—you can literally think of markets as the New York Stock Exchange) and hierarchy (a typical organization, e.g.,

a *Fortune* 500 company) on the other. Intermediate or hybrid forms of organization such as joint ventures, long-term contracts, and franchising fall somewhere in the middle between the market and hierarchy anchor points of this governance spectrum. The big differences between the two extremes have to do with the ability to adapt or coordinate action, the incentive intensity (i.e., the ability to motivate people), and the way disputes or disagreements are settled.

Because the theory adopts the premise that there is no such thing as a perfect governance solution (i.e., Williamson [e.g., 1985, 1996] emphasizes that there is little use in considering hypothetical ideals), there are trade-offs. Three dimensions constitute these trade-offs: incentives, adaptation, and dispute resolution (see Table 92.1). Note that markets enjoy high-powered incentives compared to hierarchy, hierarchy is better than markets when coordinated adaptation is required, and dispute resolution of last resort takes place in courts when market exchange is used, while hierarchy relies on internal mechanisms. From the table, it is clear that there is no form of governance that will score highly on all dimensions. Put differently, it is impossible to create a firm (hierarchy) that offers the dispute resolution and cooperative adaptation advantages as well as the high-powered incentive advantages reserved for markets. A common mistake managers make is to attempt to create just such governance structures. Williamson (1991) labels this problem "the folly of selective intervention."

The question of economic exchange thus becomes one of how to select the right governance structure. The answer to this question depends on the characteristics of the transaction. Transactions are differentiated based on three variables: asset specificity, uncertainty, and frequency (i.e., is this an exchange that you intend to repeat over time), of which asset specificity is deemed to be the critical factor. Williamson (1991) first identified three different types of asset specificity, but that list has now been expanded to six.

**Table 92.1** Comparison of Attributes of Governance Structures

| *Attributes* | *Market* | *Hybrid* | *Hierarchy* |
|---|---|---|---|
| Incentive intensity | ++ | + | 0 |
| Administrative control | 0 | + | ++ |
| Autonomous adaptation | ++ | + | 0 |
| Cooperative adaptation | 0 | + | ++ |
| Reliance on contract law | ++ | + | 0 |

SOURCE: Adapted from Williamson (1991, p. 281).

NOTES: ++ = strong; + = semi-strong; 0 = weak.

They are physical asset specificity, human asset specificity, site specificity, brand-name specificity, dedicated assets, and temporal specificity. Think of asset specificity as the degree to which an asset has become specialized for a given exchange. Highly specific assets are those that are much more valuable to a firm in the context of a given transaction and whose value is negligible outside this exchange. For example, prior to agreeing to build components for General Motors (GM), Fisher Body had tool-and-die equipment that could be tailored to stamp out body parts for any automobile manufacturer. These industrial machines had little physical asset specificity. However, once Fisher Body agreed to produce parts for GM, these machines needed to be calibrated to technical specifications unique to the GM components. Now these same tool-and-die machines would be categorized as being highly asset specific. In other words, their value to anyone other than Fisher and GM would be their scrap value.

The central insight of transaction cost alignment is that with an increase in asset specificity (as well as uncertainty and frequency), the potential hazards of relying on market-like forms of exchange increase. Think about the risks to both GM and Fisher Body if one party decides—*after* the specific investment in the tool-and-die equipment is made—to alter the terms of the contract. Fisher Body might decide to raise the price of the components. If this occurs during a period of peak demand, GM will likely be hard-pressed to find another supplier, much less find another supplier quickly enough to meet demand. Alternatively, during a period of weak demand, GM might decide to lower the price it is willing to pay. Fisher Body would find itself in a poor bargaining position and thus may likely acquiesce to this change in terms.

Recognizing ahead of time that this sort of behavior might occur is a key insight of transaction cost economics. The theory is unusual in that it specifies two important behavioral assumptions: (1) Agents are far-sighted, boundedly rational decision makers, and (2) they will behave opportunistically. Combining an appreciation for (a) unwavering differences in governance structures; (b) an ability to measure the features of transactions, particularly the level of asset specificity; and (c) these behavioral assumptions, transaction cost economics generates the following testable hypotheses, the last of which is known as the "discriminating alignment hypothesis" (Williamson, 1991):

- As asset specificity rises, contracting hazards rise.
- As contracting hazards rise, transaction costs rise.
- Thus, as asset specificity rises, we are more likely to observe hierarchical modes of governance.

Williamson (1991) articulates that managers at both Fisher Body and GM, given those stated behavioral assumptions, would recognize that the idiosyncratic (asset-specific) investments required to stamp out customized

components would lead to potential contracting hazards and thus increased transaction costs—costs that could not be remedied through pricing. Instead, as the third hypothesis above indicates, managers chose the governance structure that economizes on transaction costs. Recall that these are the costs that arise not only from the contracting hazards (e.g., the potential last-minute changes in pricing depending on external demand conditions) but also the costs of writing, monitoring, and haggling over contracts to try to prevent such opportunistic behavior. As a result, transaction cost economics predicts that the best form of governance, the form that is best able to economize on these costs, is hierarchy. And indeed, after several years of exchange, GM did adjust its governance structure accordingly and acquired Fisher Body, transforming it from an exchange partner to an embedded division within the GM organization.

As outlined in the introduction, transaction cost economics has had tremendous success in predicting the mode of organization or governance choice. According to recent reviews of the empirical literature (Boerner & Macher, 2000; Macher & Richman, 2008; Shelanski & Klein, 1995), more than 600 studies have been published, the majority of which support the primary hypotheses of the discriminating alignment argument. Consistent with theoretical predictions, high levels of asset specificity lead to an increased likelihood of firms relying on vertical integration.

Even today, the vast majority of the empirical studies that have been undertaken in transaction cost economics have tested the discriminating alignment hypothesis. But as we will highlight in the remaining sections of this chapter, this is reflective of the early timing of these papers, not the state of current empirical research. Nonetheless, to highlight the contributions of these early papers is warranted. Monteverde and Teece (1982) provided what is still seen as an archetypal study of vertical integration through a transaction cost lens. Using data on the two leading U.S. automobile producers, the authors found that, as predicted, as the level of asset specificity increased, the likelihood of vertical integration did as well. Asset specificity is measured in this paper as is still commonly done today: via surveying experts as to the degree of unique, idiosyncratic investment required to support the exchange. Generally, such surveys use multiple questions in combination as a proxy for asset specificity.

Although using surveys remains a popular measurement technique, not all studies rely on survey measures of asset specificity. For example, Joskow (1985, 1987, 1990) used the measurement of physical proximity when measuring site specificity. He finds that indeed as site specificity increases, power plants are more likely to use more hierarchical forms of exchange in accessing the coal needed for power generation.

While three variables—asset specificity, frequency, and uncertainty—are theorized to affect the choice of governance, it is interesting to note that most studies deal with either uncertainty or asset specificity, but few deal with frequency. To be sure, Williamson (1985, 1991) explains that absent asset specificity, the other two variables do not pose the same sort of governance concerns. Thus, studies that have examined uncertainty do so in conjunction with inclusions of controls for asset specificity (e.g., Anderson, 1985; Stump & Heide, 1996; Walker & Weber, 1987).

## Beyond Make-Buy: Strategic Alliances, International Business, and Performance

### Strategic Alliances

Transaction cost economics is a branch of economics best known for its contribution to the boundary of the firm literature, with particular emphasis on vertical integration (e.g., Williamson, 1975, 1985, 1996). As discussed in the preceding section, hundreds of studies (as many as 600) have largely corroborated the primary predictions of the theory (i.e., the discriminating alignment hypothesis) regarding the likelihood of integration under certain conditions. Few theoretical perspectives on the firm have been so richly supported by the empirical evidence. But transaction cost economics has made significant contributions to other research areas, contributions that may not be as well known. This section highlights the recent application of transaction cost logic to strategic alliances, international business, and performance.

The initial governance spectrum developed by Williamson was designed to illustrate the trade-offs of the two extreme forms, markets and hierarchy. Over time, researchers began to unpack the middle of the governance spectrum, what Williamson (1991, 1996) refers to as hybrid forms of governance. Given the prevalence of strategic alliances, these became the focus of much recent work on better understanding this particular hybrid form of governance.

While there had been a handful of studies that examined hybrid modes early on (e.g., Heide & John, 1990; Palay, 1984), it was not until Oxley (1999) that the field had a means for separating the features of hybrid modes. In her paper, Oxley provides evidence that these hybrid forms of organization are being used at an increasing rate, motivating interest in better understanding them. Furthermore, her work provides a spectrum to be used in conjunction with the initial governance spectrum outlined by Williamson, to guide research in determining how market-like or hierarchical a given hybrid mode is and thus what sort of trade-offs does this hybrid form offer.

Examining strategic alliances has gone beyond extensions of the discriminating alignment hypothesis as illustrated with the above research. For example, Reuer, Zollo, and Singh (2002) examine the evolution of strategic alliances, using data on biotech and pharmaceutical alliances. They specifically set out to determine whether

collaborative agreements experience significant contractual alterations in the area of control or monitoring. Interestingly, while governance does change over time, it does so more often based on the experience of the strategic alliance partners, rather than in response to changes in conditions.

## International

A review of the transaction cost empirical literature shows that it is only recently, over the past decade, that a sustained effort has been under way to expand the geographic scope of the theory. Due to both the spread of transaction cost research among scholars in institutions outside the United States as well as the research activities of U.S.-based scholars, real progress in examining issues related to international business through a transaction cost lens has been achieved. For example, work has analyzed foreign direct investment decisions, the sequence of such decisions (e.g., Delios & Henisz, 2003), the role of institutions in these decisions (e.g. Henisz, 2000), and the value created through these decisions (e.g., Reuer, 2001).

In an interesting paper that relates to both the strategic alliance and international literatures, Dyer (1996) compares the differences in governance structures used by U.S. and Japanese automobile manufacturers. He finds that U.S. firms use either end of the spectrum, relying on arm's-length contracting or vertical integration. In contrast, Japanese manufacturers rely heavily on hybrid forms of governance, especially long-term, relational contracting. The Japanese firms are able to thereby reap some of the benefits of hierarchy (e.g., they can manage higher levels of asset-specific investment) with markets (e.g., they are able to manage technical uncertainty better by switching suppliers as needs change). An important insight, however, is that it is due to the differences in the national-level institutional supports for exchange that allow the Japanese firms to be so successful. In other words, without such institutional support, U.S. firms cannot replicate the governance structure employed by their Japanese counterparts and achieve similar results.

Understanding the role of political or country-specific institutional factors in the decisions of large firms to enter new markets has been improved through the recent work of researchers working through a transaction cost lens. Henisz and colleagues (e.g., Delios & Henisz, 2000; Henisz, 2000; Henisz & Zelner, 2001) have found a method of better measuring political risk—an oft-cited rationale for the reluctance of large multinationals to enter these markets—and used it to predict not just the likelihood of entry but the likelihood of entry mode (i.e., what form of governance does the entering firm employ to offset the likely contractual hazards that may occur as a result of institutional factors). In addition to political risk, they measure the feasibility of adaptation in the institutional environment and find results that support their hypotheses regarding the use of governance mode given more favorable institutional conditions.

## Performance

As we have discussed, transaction cost economics adopts the view that firms can be thought of as bundles of transactions. And each transaction has an optimal (i.e., transaction cost economizing) way of being structured. The implication for strategy researchers is that firms that are the most optimally aligned (most efficiently structured) will outperform those that are not. The other big implication for strategy is that those firms that recognize the pitfalls of exchange will do better. These kinds of questions go beyond the basic likelihood of what governance structure is chosen by a firm. The question now becomes, if a firm does adhere to the discriminating alignment hypothesis, does it benefit as a result? Benefit, in this instance, meaning does it outperform its rivals?

Although the implications for firm performance of transaction cost economics have been of great interest to strategy scholars, the empirical demands of conducting actual tests of these potential performance hypotheses have been daunting (e.g., Masten, 1993). For example, in addition to collecting data at the microanalytic level (e.g., information on the asset specificity, uncertainty, and frequency of the transaction), as well as information on what governance mode is chosen, researchers also must collect data at the firm, industry, and/or economic levels (e.g., the performance of the firm, industry averages of performance, economic variables that might also affect performance). Note that there are also temporal differences in the nature of the data required. Researchers who study questions related to testing variations of the discriminating alignment hypothesis can use cross-sectional data. But strategy researchers interested in tracking the impact of firm decisions on performance require longitudinal data. This presents enormous empirical obstacles that must be overcome. As a result, it is only recently that progress has been made in this area.

Early studies in this area measured performance as something other than profitability. Nickerson and Bigelow (2008) summarize much of this initial performance research. They describe the finding that the transfer of technological knowledge is diminished as firms must rely on market governance given relatively high levels of asset specificity (Poppo & Zenger, 1998) while Leiblein, Reuer, and Dalsace (2002) find that the alignment of governance choice and contracting hazards ultimately improves technological performance. Using data on R&D alliances, Sampson (2004) finds that the alignment of transactions according to transaction cost predictions conferred collaborative benefits not found in transactions organized otherwise. Transaction cost economics has begun to amass a body of research that indicates that firms that achieve an efficient alignment enjoy performance benefits. However,

none of these studies estimates economic performance in terms of profitability or cost savings.

The first study to provide estimates of economic performance at a transaction level was Masten, Meehan, and Snyder (1991). Their study found that cost savings in shipbuilding were achieved in conjunction with organizing governance structures as predicted by transaction cost theory. In developing their empirical estimates, they also used econometric methods that statistically remedied the endogeneity problem inherent in doing comparative analysis that accounts for the selection of discrete organizational forms. While this study offered a breakthrough in empirical transaction cost research by estimating cost savings, it still did not achieve the long-sought-after goal of estimating economic profits at the transaction level.

To date, the first and only study to provide estimates of profitability at a transaction level is Mayer and Nickerson (2005). Analyzing the contracts of an information technology company, the authors first test the discriminating alignment hypothesis. Then, using a two-stage switching regression model, they show that projects aligned according to the discriminating alignment hypothesis are, on average, more profitable than misaligned projects. Note that this is another econometric solution to the endogeneity problem endemic to transaction cost research that aims to study performance.

While measuring the impact of profitability and cost savings remains rare, researchers in this area have demonstrated creativity in adopting alternate proxies for financial performance. For example, Silverman, Nickerson, and Freeman (1997); Nickerson and Silverman (2003a, 2003b); and Argyres and Bigelow (2007) employ the duration and survivability of firms as substitutes for financial performance. The crux of the argument is that in competitive markets, firms that survive longer than their rivals must have greater profitability and/or access to financial resources. These studies find that consistent with transaction cost predictions, firms that use governance structures aligned with transaction characteristics are more likely to live longer and are less likely to fail than firms that do not organize efficiently.

## Emerging Areas of Research: Opportunities in the Areas of Contracts, Technological Evolution, and Industry Dynamics

While there remain many opportunities to build on the areas of strategic alliances (and other hybrid governance structures), international business, and performance, at least two areas of research in transaction cost economics have emerged only recently: the study of contractual features and the role of technological evolution and industry dynamics.

### Contracts

In a study designed to better understand the nature of contractual learning processes over time, Argyres, Bercovitz, and Mayer (2007) examine two categories of contractual provisions that are used to manage uncertainty and asset specificity. Using data from an information technology provider over an 18-year period, the authors find that certain contractual clauses related to contingency planning and task description did change over time. They suggest that time-dependent reputation advantages may have played a role, providing an effective and efficient safeguard against opportunistic behavior.

Other studies in the contracting area have focused on the role of trust not just as a complement but as a substitute for other contractual provisions. Gulati and Nickerson (2008) argue that trust acts as a shift parameter that lowers governance costs for all modes of governance whenever exchange hazards are present and thus enhances performance regardless of the mode of governance chosen. This lowering of governance cost arises because trust, which is less formal than either contracts or ownership, facilitates adaptation—exchange partners are more likely to avoid disputes or resolve them quickly when trust is present (Gulati, Lawrence, & Puranam, 2005). Gulati and Nickerson's (2008) theory also suggests that trust can lead to a substitution of less formal for more formal modes of governance because governance cost-reducing benefits of trust are greater for market than for hybrid and greater for hybrid than for hierarchy. These differences arise because trust proves a less useful safeguard when formal mechanisms such as contracts and ownership are used. The result of this differential impact is that the market mode of governance, with the addition of preexisting trust, may be used over a broader range of exchange hazards than can market sans trust, which in turn offers lower governance costs and enhances exchange performance. Also, a hybrid with preexisting trust can substitute over some range of exchange hazards for hierarchy, which enhances exchange performance. Drawing on a sample of 222 sourcing arrangements for components from two assemblers in the automobile industry, Gulati and Nickerson (2008) find broad support for both substitutive and complementary affects of interorganizational trust on qualitative measures of perceived exchange performance.

Other additional areas of recent interest include understanding the role of contractual features (e.g., Mellewigt, Madhok, & Weibel, 2007), the range of contractual provisions (e.g., Oxley & Wada 2009), and the development of contracting capabilities (e.g., Argyres & Mayer, 2004). Again, as in the performance area, the current obstacle to more research is accessing suitable data, not lack of interest.

### Technical Evolution and Industry Dynamics

As stated earlier, in the section on performance, the ability to withstand selection pressures and survive longer than rivals as a result of adhering to the discriminating alignment hypothesis has been used in a handful of studies. A branch of this research is now focused on linking the selection environment to the evolution of the industry. Furthermore,

this focus highlights the fact that selection environments change over time. Unfortunately, the presumption in transaction cost economics is that this selection pressure hinges on alignment and is made without a discussion of how factors such as the industry life cycle and degree of technological evolution might drive changes in this selection pressure.

Thus, important questions remain regarding the time required before selection pressures drive less efficient (i.e., transaction cost-economizing) firms from the industry as well as how the force of these selection pressures changes over time. Williamson (1985, p. 23) briefly suggests that efficient transaction cost economizing might occur over 5 to 10 years, though the timescale required to achieve efficient organization is rarely addressed in empirical studies. One exception is Nickerson and Silverman (2003a), who found that institutional constraints on firms in the U.S. trucking industry slowed their efforts to economize on transaction costs after deregulation.

Industry life cycle theories (e.g., Abernathy & Utterback, 1978; Klepper, 1996; Klepper & Miller, 1995) are of particular interest because they postulate general patterns in the waxing and waning of selection forces that may be and have been made subject to empirical confirmation. Thus, integrating theories of industry and technological evolution, as well as competitive intensity with transaction cost economics, may help address the need for estimating the selection pressures in operation in a given empirical context.

In the first study of this kind, Argyres and Bigelow (2007) integrate industry life cycle theory with transaction cost economics to examine the impact of organizational choice and firm survival over time. They find that firms that misalign transactions face increased risk of failure. However, this risk is mitigated by environmental selection pressures. Research on industry life cycles demonstrates that competitive pressures are more severe during the shakeout stage, which occurs later in the industry life cycle, than at other stages. Transaction cost theory, on the other hand, assumes generally competitive markets and does not address the industry life cycle. It therefore implies that transaction cost economizing is a superior firm strategy regardless of the stage of the life cycle. The authors find that, indeed, adhering to the discriminating alignment hypothesis is not a superior strategy under all time periods. Analyzing data from the early U.S. auto industry, they find that while transaction cost economizing did not have a significant impact on firm survival during the pre-shakeout stage, it did have a significant positive impact on survival during the shakeout stage. This suggests that applications of transaction cost theory, which assume uniformly severe selection pressures across the industry life cycle, could be misleading. It also suggests that theories of the industry life cycle could usefully take transaction costs into account along with production costs in their analyses of competition over the life cycle.

## Discussion and Conclusion

The purpose of this chapter has been to familiarize the reader with the key insights and contributions of transaction cost economics as well as highlight the many potential applications of the theory. As chief architect, Oliver Williamson has combined ideas from economics (e.g., Coase, 1937), law (e.g., Macaulay, 1963), and organization theory (e.g., Simon, 1957, 1959, 1962). Over the past three decades, he has refined his thinking, and other researchers have extended the empirical and theoretical reach of the theory. With fair warning to readers, it should be noted that the author still remembers reading the quote that introduced this chapter as a graduate student in a seminar taught by Williamson himself and has indeed seen her worldview of organizations irreversibly altered as a result. Thus, although every effort was made to be even-handed in discussing the current state of viewing economic problems of organization through a transaction cost lens, the reader should be aware that there may be some inadvertent bias.

Recall that in brief, the transaction cost approach works as follows: The decision makers within a firm begin by (a) understanding the trade-offs between different modes of governance. This relies on thinking of governance as falling on a spectrum with markets (arm's-length contracts) on one end and hierarchy (a typical organization, e.g., Monsanto) on the other. Hybrid forms of organization such as joint ventures, long-term contracts, and franchising fall somewhere in the middle. The big differences between the two extremes have to do with the ability to adapt (coordinate), the incentive intensity (e.g., a manager's ability to motivate people), and the way disputes are adjudicated. The next step has to do with (b) understanding the characteristics of the transaction. The three characteristics we think are critical are uncertainty, frequency (i.e., is this an exchange that the firm intends to repeat over time), and the primary driver of governance choice—asset specificity. Asset specificity is the degree to which an asset has become specialized for a given exchange. The next step is to (c) outline the contracting hazards that may arise, depending on the nature of the transaction and the features of potential governance solutions. Finally, (d) choose the governance structure that economizes on transaction costs. These are the costs that arise not only from the contracting hazards but also the costs of writing, monitoring, and haggling over contracts.

So, the immediate implications of transaction cost economics are that (a) as asset specificity increases, contracting hazards intensify; (b) as asset specificity increases and/or contracting hazards intensify, managers are likely to find that a hierarchical mode of governance is more efficient; and (c) in application, no form of governance is perfect. There are always trade-offs. The role of managers, then, is to assess those trade-offs in a systematic way and anticipate the shifts in trade-offs that will occur over time as the nature of the underlying transaction changes over time.

One additional advantage of using a transaction cost economics approach: The theory forces researchers to be precise about what questions need to be asked, what the objectives of the exchange are, and what trade-offs can be tolerated.

Furthermore, note that even the oldest research puzzles are fair game for renewed discussion. For example, a recent series of papers revisiting Fisher Body–GM (e.g., Coase, 2000; Goldberg, 2008; Klein, 1988, 2007, 2008) suggests that debates regarding governance choices, contracting hazards, and transaction costs are far from settled. This does not mean that the theory itself is not valid. Quite the contrary, as the various literature reviews have documented. Instead, this renewed debate suggests that transaction cost theory remains a robust and active area of economics, law, and organization theory. It is hoped that the student of transaction cost economics will be inspired by this and, perhaps, contribute to the conversation.

*Author's Note:* I thank Nick Argyres, Jackson Nickerson, and Todd Zenger for helpful discussions. Nonetheless, this manuscript reflects my own biases, and thus any inadvertent errors or omissions are solely the responsibility of this author.

## References and Further Readings

Abernathy, W. J., & Utterback, J. (1978, June–July). Patterns of industrial innovation. *Technology Review,* pp. 40–47.

Anderson, E. (1985). The salesperson as outside agent or employee: A transaction cost analysis. *Marketing Science, 4,* 234.

Argyres, N., & Bigelow, L. (2007). Does transaction misalignment matter for firm survival across all stages of the industry lifecycle? *Management Science, 53,* 1332–1345.

Argyres, N., & Mayer, K. (2004). Learning to contract: Evidence from the personal computer industry. *Organization Science, 15,* 394–410.

Argyres, N., & Mayer, K. (2007). Contract design as a firm capability: An integration of learning and transaction cost perspectives. *Academy of Management Review, 32,* 1060–1077.

Argyres, N. S., Bercovitz, J., & Mayer, K. J. (2007). Complementarity and evolution of contractual provisions: An empirical study of IT services contracts. *Organization Science, 18,* 3–19.

Boerner, C., & Macher, J. (2000). *Transaction cost economics: An assessment of empirical research in the social sciences* (Working paper). Washington, DC: Georgetown University.

Coase, R. H. (1937). The nature of the firm. *Economica, 4,* 386–405.

Coase, R. H. (1972). Industrial organization: A proposal for research. In V. R. Fuchs (Ed.), *Policy issues and research opportunities in industrial organization* (pp. 59–73). New York: National Bureau of Economic Research.

Coase, R. H. (2000). The acquisition of Fisher Body by General Motors. *Journal of Law & Economics, 43,* 15–31.

Delios, A., & Henisz, W. J. (2000). Japanese firms' investment strategies in emerging economies. *Academy of Management Journal, 43,* 305–323.

Delios, A., & Henisz, W. J. (2003). Political hazards, experience, and sequential entry strategies: The international expansion of Japanese firms 1980–1998. *Strategic Management Journal, 24,* 1153–1164.

Dyer, J. (1996). Does governance matter? Keiretsu alliances and asset specificity as sources of Japanese competitive advantage. *Organization Science, 7,* 649–666.

Goldberg, V. P. (2008). Lawyers asleep at the wheel? The GM–Fisher Body contract. *Industrial & Corporate Change, 17,* 1071–1084.

Grossman, S. J., & Hart, O. (1986). The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy, 94,* 691–719.

Gulati, R., Lawrence, P. R., & Puranam, P. (2005). Adaptation in vertical relationships: Beyond incentive conflict. *Strategic Management Journal, 26,* 415–440.

Gulati, R., & Nickerson, J. A. (2008). Interorganizational trust, governance choice, and exchange performance. *Organization Science, 19,* 688–708.

Hart, O., & Moore, J. (1990). Property rights and the nature of the firm. *Journal of Political Economy, 98,* 1119–1159.

Heide, J., & John, G. (1990). Alliances in industrial purchasing: The determinants of joint action in buyer-supplier relationships. *Journal of Marketing Research, 27,* 24–36.

Henisz, W. J. (2000). The institutional environment for multinational investment. *Journal of Law, Economics & Organization, 16,* 334–364.

Henisz, W. J., & Zelner, B. A. (2001). The institutional environment for telecommunications investment. *Journal of Economics & Management Strategy, 10,* 123–147.

Henisz, W. J., & Zelner, B. A. (2005). Legitimacy, interest group pressures, and change in emergent institutions: The case of foreign investors and host country governments. *Academy of Management Review, 30,* 361–382.

Joskow, P. (1985). Vertical integration and long-term contracts: The case of coal-burning electric generating plants. *Journal of Law, Economics, and Organization, 1,* 33–80.

Joskow, P. (1987). Contract duration and relation specific investments: Empirical evidence from coal markets. *American Economic Review, 17,* 168–185.

Joskow, P. (1990). The performance of long-term contracts: Further evidence from the coal markets. *Rand Journal of Economics, 21,* 251–274.

Klein, B. (1988). Vertical integration as organizational ownership: The Fisher Body–General Motors relationship revisited. *Journal of Law, Economics & Organization, 4,* 199–214.

Klein, B. (2007). The economic lessons of Fisher Body–General Motors. *International Journal of the Economics of Business, 14,* 1–36.

Klein, B. (2008). The enforceability of the GM–Fisher Body contract: Comment on Goldberg. *Industrial & Corporate Change, 17,* 1085–1096.

Klein, B., Crawford, V., & Alchian, A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics, 21,* 297–326.

Klepper, S. (1996). Entry, exit, growth, and innovation over the product life cycle. *American Economic Review, 86,* 562–583.

Klepper, S., & Miller, J. (1995). Entry, exit, and shakeouts in the United States in new manufactured products. *International Journal of Industrial Organization, 13,* 567–592.

Leiblein, M. J., Reuer, J. J., & Dalsace, F. (2002). Do make or buy decisions matter? The influence of organizational governance on technological performance. *Strategic Management Journal, 23,* 817–834.

Macaulay, S. (1963). Non-contractual relations in business. *American Sociological Review, 28,* 55–70.

Macher, J. (2006). Technological development and the boundaries of the firm: A knowledge-based examination in semiconductor manufacturing. *Management Science, 52,* 826–843.

Macher, J., & Richman, B. (2008). Transaction cost economics: An assessment of empirical research in the social sciences. *Business and Politics, 10*(1), 1–63.

Masten, S. (1993). Transaction costs, mistakes and performance: Assessing the importance of governance. *Managerial and Decision Economics, 14,* 119–130.

Masten, S., Meehan, M., & Snyder, E. (1991). The costs of organization. *Journal of Law, Economics and Organization, 7,* 1–25.

Mayer, K. J., & Nickerson, J. A. (2005). Antecedents and performance implications of contracting for knowledge workers: Evidence from information technology services. *Organization Science, 16,* 225–242.

Mellewigt, T., Madhok, A., & Weibel, A. (2007). Trust and formal contracts in interorganizational relationships: Substitutes and complements. *Managerial & Decision Economics, 28,* 833–847.

Monteverde, K., & Teece, D. (1982). Supplier switching costs and vertical integration in the automobile industry. *Bell Journal of Economics, 13,* 206–213.

Nickerson, J. A., & Bigelow, L. S. (2008). New institutional economics, organization, and strategy. In J.-M. Glachant & E. Brousseau (Eds.), *New institutional economics: A guidebook* (pp. 183–208). Cambridge, UK: Cambridge University Press.

Nickerson, J., & Silverman, B. (2003a). Why firms want to organize efficiently and what keeps them from doing so: Evidence from the for-hire trucking industry. *Administrative Science Quarterly, 3,* 433–465.

Nickerson, J. A., & Silverman, B. S. (2003b). Why firms want to organize efficiently and what keeps them from doing so: Inappropriate governance, performance, and adaptation in a deregulated industry. *Administrative Science Quarterly, 48,* 433–465.

Oxley, J. (1999). Institutional environment and the mechanisms of governance: The impact of intellectual property. *Journal of Economic Behavior & Organization, 38,* 283–310.

Oxley, J., & Wada, T. (2009). Alliance structure and the scope of knowledge transfer: Evidence from US-Japan agreements. *Management Science, 55,* 635–649.

Palay, T. (1984). Comparative institutional economics: The governance of rail freight contracting. *Journal of Legal Studies, 13,* 265–288.

Poppo, L., & Zenger, T. (1998). Testing alternative theories of the firm: Transaction cost, knowledge-based, and measurement explanations for make-or-buy decisions in IT services. *Strategic Management Journal, 19,* 853–877.

Reuer, J. J. (2001). From hybrids to hierarchies: Shareholder wealth effects of joint venture partner buyouts. *Strategic Management Journal, 22,* 27–45.

Reuer, J. J., Zollo, M., & Singh, H. (2002). Post-formation dynamics in strategic alliances. *Strategic Management Journal, 23,* 135–152.

Riordan, M., & Williamson, O. E. (1985). Asset specificity and economic organization. *International Journal of Industrial Organization, 3,* 365–378.

Sampson, R. C. (2004). The cost of misaligned governance in R&D alliances. *Journal of Law, Economics & Organization, 20,* 484–526.

Shelanski, H., & Klein, P. (1995). Empirical research in transaction cost economics. *Journal of Law, Economics and Organization, 11,* 335–361.

Silverman, B., Nickerson, J. A., & Freeman, J. (1997). Profitability, transactional alignment, and organizational mortality in the U.S. trucking industry. *Strategic Management Journal, 18,* 31–52.

Simon, H. (1957). *Administrative behavior* (2nd ed.). New York: Macmillan.

Simon, H. (1959). Theories of decision-making in economics and behavioral science. *American Economic Review, 49,* 253–284.

Simon, H. (1962). New developments in the theory of the firm. *American Economic Review, 52,* 1–15.

Stump, R. L., & Heide, J. B. (1996). Controlling supplier opportunism in industrial relationships. *Journal of Marketing Research, 33,* 431–441.

Suarez, F., & Utterback, J. (1995). Dominant designs and the survival of firms. *Strategic Management Journal, 16,* 415–430.

Walker, G., & Weber, D. (1987). Supplier competition, uncertainty, and make-or-buy decisions. *Academy of Management Journal, 30,* 589–596.

Williamson, O. E. (1975). *Markets and hierarchies.* New York: Free Press.

Williamson, O. E. (1985). *The economic institutions of capitalism.* New York: Free Press.

Williamson, O. E. (1991). Comparative economic organization: The analysis of discrete structural alternatives. *Administrative Science Quarterly, 36,* 269–296.

Williamson, O. E. (1996). *The mechanisms of governance.* New York: Oxford University Press.

# INDEX

Main topics and their page numbers are in **bold.** Page numbers referring to figures and tables are followed by (figure) and (table).