Public Administration and Public Policy/125

Handbook of Public Policy Analysis

Theory, Politics, and Methods

Edited by
Frank Fischer
Gerald J. Miller
Mara S. Sidney

CRC Press
Taylor & Francis Group

Handbook of Public Policy Analysis

Theory, Politics, and Methods

PUBLIC ADMINISTRATION AND PUBLIC POLICY

A Comprehensive Publication Program

Executive Editor

JACK RABIN
Professor of Public Administration and Public Policy
School of Public Affairs
The Capital College
The Pennsylvania State University—Harrisburg
Middletown, Pennsylvania

Assistant to the Executive Editor
T. Aaron Wachhaus, Jr.

1. *Public Administration as a Developing Discipline,* Robert T. Golembiewski
2. *Comparative National Policies on Health Care,* Milton I. Roemer, M.D.
3. *Exclusionary Injustice: The Problem of Illegally Obtained Evidence,* Steven R. Schlesinger
5. *Organization Development in Public Administration,* edited by Robert T. Golembiewski and William B. Eddy
7. *Approaches to Planned Change,* Robert T. Golembiewski
8. *Program Evaluation at HEW,* edited by James G. Abert
9. *The States and the Metropolis,* Patricia S. Florestano and Vincent L. Marando
11. *Changing Bureaucracies: Understanding the Organization before Selecting the Approach,* William A. Medina
12. *Handbook on Public Budgeting and Financial Management,* edited by Jack Rabin and Thomas D. Lynch
15. *Handbook on Public Personnel Administration and Labor Relations,* edited by Jack Rabin, Thomas Vocino, W. Bartley Hildreth, and Gerald J. Miller
19. *Handbook of Organization Management,* edited by William B. Eddy
22. *Politics and Administration: Woodrow Wilson and American Public Administration,* edited by Jack Rabin and James S. Bowman
23. *Making and Managing Policy: Formulation, Analysis, Evaluation,* edited by G. Ronald Gilbert
25. *Decision Making in the Public Sector,* edited by Lloyd G. Nigro
26. *Managing Administration,* edited by Jack Rabin, Samuel Humes, and Brian S. Morgan
27. *Public Personnel Update,* edited by Michael Cohen and Robert T. Golembiewski
28. *State and Local Government Administration,* edited by Jack Rabin and Don Dodd
29. *Public Administration: A Bibliographic Guide to the Literature,* Howard E. McCurdy
31. *Handbook of Information Resource Management,* edited by Jack Rabin and Edward M. Jackowski
32. *Public Administration in Developed Democracies: A Comparative Study,* edited by Donald C. Rowat
33. *The Politics of Terrorism: Third Edition,* edited by Michael Stohl

34. *Handbook on Human Services Administration,* edited by Jack Rabin and Marcia B. Steinhauer
36. *Ethics for Bureaucrats: An Essay on Law and Values, Second Edition,* John A. Rohr
37. *The Guide to the Foundations of Public Administration,* Daniel W. Martin
39. *Terrorism and Emergency Management: Policy and Administration,* William L. Waugh, Jr.
40. *Organizational Behavior and Public Management: Second Edition,* Michael L. Vasu, Debra W. Stewart, and G. David Garson
43. *Government Financial Management Theory,* Gerald J. Miller
46. *Handbook of Public Budgeting*, edited by Jack Rabin
49. *Handbook of Court Administration and Management*, edited by Steven W. Hays and Cole Blease Graham, Jr.
50. *Handbook of Comparative Public Budgeting and Financial Management*, edited by Thomas D. Lynch and Lawrence L. Martin
53. *Encyclopedia of Policy Studies: Second Edition,* edited by Stuart S. Nagel
54. *Handbook of Regulation and Administrative Law,* edited by David H. Rosenbloom and Richard D. Schwartz
55. *Handbook of Bureaucracy,* edited by Ali Farazmand
56. *Handbook of Public Sector Labor Relations*, edited by Jack Rabin, Thomas Vocino, W. Bartley Hildreth, and Gerald J. Miller
57. *Practical Public Management*, Robert T. Golembiewski
58. *Handbook of Public Personnel Administration*, edited by Jack Rabin, Thomas Vocino, W. Bartley Hildreth, and Gerald J. Miller
60. *Handbook of Debt Management*, edited by Gerald J. Miller
61. *Public Administration and Law: Second Edition*, David H. Rosenbloom and Rosemary O'Leary
62. *Handbook of Local Government Administration*, edited by John J. Gargan
63. *Handbook of Administrative Communication*, edited by James L. Garnett and Alexander Kouzmin
64. *Public Budgeting and Finance: Fourth Edition,* edited by Robert T. Golembiewski and Jack Rabin
67. *Handbook of Public Finance*, edited by Fred Thompson and Mark T. Green
68. *Organizational Behavior and Public Management: Third Edition,* Michael L. Vasu, Debra W. Stewart, and G. David Garson
69. *Handbook of Economic Development*, edited by Kuotsai Tom Liou
70. *Handbook of Health Administration and Policy*, edited by Anne Osborne Kilpatrick and James A. Johnson
71. *Handbook of Research Methods in Public Administration*, edited by Gerald J. Miller and Marcia L. Whicker
72. *Handbook on Taxation*, edited by W. Bartley Hildreth and James A. Richardson
73. *Handbook of Comparative Public Administration in the Asia-Pacific Basin*, edited by Hoi-kwok Wong and Hon S. Chan
74. *Handbook of Global Environmental Policy and Administration,* edited by Dennis L. Soden and Brent S. Steel
75. *Handbook of State Government Administration,* edited by John J. Gargan
76. *Handbook of Global Legal Policy,* edited by Stuart S. Nagel
78. *Handbook of Global Economic Policy,* edited by Stuart S. Nagel
79. *Handbook of Strategic Management: Second Edition,* edited by Jack Rabin, Gerald J. Miller, and W. Bartley Hildreth
80. *Handbook of Global International Policy,* edited by Stuart S. Nagel

81. *Handbook of Organizational Consultation: Second Edition,* edited by Robert T. Golembiewski
82. *Handbook of Global Political Policy,* edited by Stuart S. Nagel
83. *Handbook of Global Technology Policy,* edited by Stuart S. Nagel
84. *Handbook of Criminal Justice Administration*, edited by M. A. DuPont-Morales, Michael K. Hooper, and Judy H. Schmidt
85. *Labor Relations in the Public Sector: Third Edition*, edited by Richard C. Kearney
86. *Handbook of Administrative Ethics: Second Edition,* edited by Terry L. Cooper
87. *Handbook of Organizational Behavior: Second Edition*, edited by Robert T. Golembiewski
88. *Handbook of Global Social Policy,* edited by Stuart S. Nagel and Amy Robb
89. *Public Administration: A Comparative Perspective, Sixth Edition,* Ferrel Heady
90. *Handbook of Public Quality Management*, edited by Ronald J. Stupak and Peter M. Leitner
91. *Handbook of Public Management Practice and Reform*, edited by Kuotsai Tom Liou
92. *Personnel Management in Government: Politics and Process, Fifth Edition,* Jay M. Shafritz, Norma M. Riccucci, David H. Rosenbloom, Katherine C. Naff, and Albert C. Hyde
93. *Handbook of Crisis and Emergency Management*, edited by Ali Farazmand
94. *Handbook of Comparative and Development Public Administration: Second Edition,* edited by Ali Farazmand
95. *Financial Planning and Management in Public Organizations*, Alan Walter Steiss and Emeka O. Cyprian Nwagwu
96. *Handbook of International Health Care Systems,* edited by Khi V. Thai, Edward T. Wimberley, and Sharon M. McManus
97. *Handbook of Monetary Policy,* edited by Jack Rabin and Glenn L. Stevens
98. *Handbook of Fiscal Policy,* edited by Jack Rabin and Glenn L. Stevens
99. *Public Administration: An Interdisciplinary Critical Analysis,* edited by Eran Vigoda
100. *Ironies in Organizational Development: Second Edition, Revised and Expanded,* edited by Robert T. Golembiewski
101. *Science and Technology of Terrorism and Counterterrorism*, edited by Tushar K. Ghosh, Mark A. Prelas, Dabir S. Viswanath, and Sudarshan K. Loyalka
102. *Strategic Management for Public and Nonprofit Organizations*, Alan Walter Steiss
103. *Case Studies in Public Budgeting and Financial Management: Second Edition,* edited by Aman Khan and W. Bartley Hildreth
104. *Handbook of Conflict Management,* edited by William J. Pammer, Jr. and Jerri Killian
105. *Chaos Organization and Disaster Management,* Alan Kirschenbaum
106. *Handbook of Gay, Lesbian, Bisexual, and Transgender Administration and Policy,* edited by Wallace Swan
107. *Public Productivity Handbook: Second Edition,* edited by Marc Holzer
108. *Handbook of Developmental Policy Studies,* edited by Gedeon M. Mudacumura, Desta Mebratu and M. Shamsul Haque
109. *Bioterrorism in Medical and Healthcare Administration,* Laure Paquette
110. *International Public Policy and Management: Policy Learning Beyond Regional, Cultural, and Political Boundaries,* edited by David Levi-Faur and Eran Vigoda-Gadot
111. *Handbook of Public Information Systems, Second Edition,* edited by G. David Garson
112. *Handbook of Public Sector Economics,* edited by Donijo Robbins

113. *Handbook of Public Administration and Policy in the European Union,* edited by M. Peter van der Hoek
114. *Nonproliferation Issues for Weapons of Mass Destruction,* Mark A. Prelas and Michael S. Peck
115. *Common Ground, Common Future: Moral Agency in Public Administration, Professions, and Citizenship*, Charles Garofalo and Dean Geuras
116. *Handbook of Organization Theory and Management: The Philosophical Approach, Second Edition,* edited by Thomas D. Lynch and Peter L. Cruise
117. *International Development Governance,* edited by Ahmed Shafiqul Huque and Habib Zafarullah
118. *Sustainable Development Policy and Administration*, edited by Gedeon M. Mudacumura, Desta Mebratu, and M. Shamsul Haque
119. *Public Financial Management,* edited by Howard A. Frank
120. *Handbook of Juvenile Justice: Theory and Practice,* edited by Barbara Sims and Pamela Preston
121. *Emerging Infectious Diseases and the Threat to Occupational Health in the U.S. and Canada,* edited by William Charney
122. *Handbook of Technology Management in Public Administration,* edited by David Greisler and Ronald J. Stupak
123. *Handbook of Decision Making,* edited by Göktuğ Morçöl
124. *Handbook of Public Administration, Third Edition,* edited by Jack Rabin, W. Bartley Hildreth, and Gerald J. Miller
125. *Handbook of Public Policy Analysis: Theory, Politics, and Methods,* edited by Frank Fischer, Gerald J. Miller, and Mara S. Sidney
126. *Elements of Effective Governance: Measurement, Accountability and Participation,* Kathe Callahan

Available Electronically

Principles and Practices of Public Administration, edited by Jack Rabin, Robert F. Munzenrider, and Sherrie M. Bartell

Handbook of Public Policy Analysis

Theory, Politics, and Methods

Edited by

Frank Fischer
Rutgers University
Newark, New Jersey, U.S.A.

Gerald J. Miller
Rutgers University
Newark, New Jersey, U.S.A.

Mara S. Sidney
Rutgers University
Newark, New Jersey, U.S.A.

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4

International Standard Book Number-10: 1-57444-561-8 (Hardcover)
International Standard Book Number-13: 978-1-57444-561-9 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Handbook of public policy analysis: theory, politics, and methods / edited by Frank Fischer, Gerald J. Miller, and Mara S. Sidney.
p. cm. -- (Public administration and public policy ; 125)
Includes bibliographical references and index.
ISBN-13: 978-1-57444-561-9 (alk. paper)
ISBN-10: 1-57444-561-8 (alk. paper)
1. Policy sciences--Handbooks, manuals, etc. 2. Public administration--Handbooks, manuals, etc. I. Fischer, Frank, 1942- II. Miller, Gerald. III. Sidney, Mara S., 1964- IV. Title. V. Series.

H97.H3583 2007
352.3'4--dc22 2006031906

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contributors

Clinton J. Andrews is an associate professor in the Edward J. Bloustein School of Planning and Public Policy at Rutgers University and director of the Urban Planning Program. He has published widely on energy and environmental management and policy, and his most recent book is *Humble Analysis*.

Thomas A. Birkland directs the Center for Policy Research, State University of New York at Albany, where he is also a professor. He is the author of *After Disaster: Agenda Setting, Public Policy, and Focusing Events.*

Susan E. Clarke is professor of political science at the University of Colorado at Boulder. She teaches a graduate seminar on context-sensitive research methods. She is an editor of *Urban Affairs Review*. Her most recent book is *The Work of Cities* (co-authored with Gary Gaile).

Caroline Danielson is a policy analyst at the Public Policy Institute of California, in San Francisco. She earned her doctorate in political science from the University of Michigan, Ann Arbor.

Peter deLeon earned his Ph.D. from the Rand Graduate School. Dr. deLeon is the author of *Democracy and the Policy Sciences* as well as *Advice and Consent.*

Tansu Demir, PhD, is assistant professor in the Department of Public Administration at the University of Central Florida. He received his Ph.D. in public administration from Florida Atlantic University in 2005.

Frank Fischer is professor of political science and member of the faculty of the Edward J. Bloustein School of Planning and Public Policy at Rutgers University. His recent publications include *Reframing Public Policy: Discursive Politics and Deliberative Practice,* and *Citizens, Experts, and the Environment: The Politics of Local Knowledge.*

John Forester is professor of city and regional planning at Cornell University. His best known work includes *The Deliberative Practitioner, Planning in the Face of Power* (University of California Press, 1989), and *The Argumentative Turn in Policy Analysis and Planning* (co-edited with Frank Fischer).

Jan-Eric Furubo, an evaluator and has been at the National Audit Office in Sweden, is the author of many articles and publications in the field of decision making, and was co-editor of the *International Atlas of Evaluation* (2002). He is president of the Swedish Evaluation Society.

Yaakov Garb is a lecturer at the Jacob Blaustein Institutes for Desert Research at the Ben-Gurion University of the Negev, and a visiting assistant professor in the Global Environmental Program at the Watson Institute for International Studies, Brown University. He has worked on a range of environmental and urban issues internationally, often drawing on perspectives from Science and Technology Studies (STS). He has recently completed essays on the "construction of inevitability" in megaprojects, on changing retail travel patterns in Central Europe, and on the politics of mobility in Israel and Palestine.

Herbert Gottweis is director at the Department of Political Science of the University of Vienna. His publications include *Governing Molecules*: *The Discursive Politics of Genetic Engineering in Europe and in the United States*.

Steven Griggs is lecturer in public policy at the Institute of Local Government Studies at the University of Birmingham in the UK. His current research centres on discourses of community protest campaigns against the expansion of airports in the UK.

John Grin is a professor of policy science at the Department of Political Science at the University of Amsterdam. He is also Director of the Amsterdam School for Social Science Research, and co-director of the Dutch Knowledge Network on System Innovations, a research program on fundamental transitions to a sustainable society.

Hubert Heinelt is professor for public administration, public policy and urban and regional research at Darmstadt University of Technology. He is a member of the executive committee of the European Urban Research Association and the Standing Group on Urban Research of the German Political Science Association.

Robert Hoppe is a professor in the Faculty of Business, Public Administration, and Technology (BBT), University of Twente, Netherlands. He is chair of Policy and Knowledge and editor-in-chief of *Beleidswetenschap*. His key research interests are in methods of policy analysis and science/policy boundary work.

Helen Ingram is Warmington Endowed Chair of Social Ecology at the University of California at Irvine. She has joint appointments in the Departments of Planning, Policy and Design, Political Science, and Criminology, Law and Society.

Werner Jann holds the chair for Political Science, Administration and Organisation at the University of Potsdam, Germany. He was associate professor at the Postgraduate School of Administrative Sciences Speyer, and has been research fellow at the University of California, Berkeley.

Patrick Kenis is professor at Tilburg University, the Netherlands, where he is also head of Department Organisation Studies. He earned his Ph.D. in social and political sciences from the European University Institute in Florence, Italy.

David Laws is principal research scientist and lecturer in the Department of Urban Studies and Planning at the Massachusetts Institute of Technology. His recent publications include *Reframing Regulation: Changing Forms of Law and Practice in U.S. Environmental Policy*, and *The Practice of Innovation: Institutions, Policy, and Technology Development.*

Anne Loeber is a post-doctoral researcher and lecturer in public policy at the Department of Political Science at the University of Amsterdam, the Netherlands. She is also a member of the Technology Assessment steering committee, an independent advisory body to the Dutch Ministry of Agriculture, Nature Managment, and Fisheries.

Martin Lodge is lecturer in political science and public policy at the Department of Government and the ESRC Centre for Analysis of Risk and Regulation, London School of Economics and Political Science. His key research interests are in comparative executive government, in particular in the area of regulation.

Miriam Manon is a graduate of the University of Massachusetts, Amherst's Commonwealth Honors College, where she earned an interdisciplinary B.A. in social justice and the environment. She completed a semester at the Arava Institute for Environmental Studies in Israel and plans to continue her studies on the interface of environmental and social issues.

Kuldeep Mathur recently retired as academic director at the Centre for the Study of Law and Governance, and professor at the Centre for Political Studies, Jawaharal Nehru University (JNU), New Delhi, India. He was formerly rector at JNU and director of India's National Institute of Education Planning and Administration.

Navdeep Mathur is research fellow at the Institute of Local Government Studies, School of Public Policy, University of Birmingham, UK. He is also forums editor of the *Journal of Critical Policy Analysis.*

Igor Mayer is an associate professor in the faculty of Technology, Policy and Management at Delft University of Technology, the Netherlands. He is also the director of the Delft-Rotterdam Centre for Process Management and Simulation.

Gerald J. Miller is professor of public administration at Rutgers University, where he teaches government and nonprofit budgeting and financial management. He has published numerous books and research articles, including *The Handbook of Debt Management* and *Government Financial Management Theory.*

Hugh T. Miller is professor of public administration and director of the School of Public Administration at Florida Atlantic University. His most recent books are *Postmodern Public Administration: Revised Edition,* with the late Charles J. Fox and *Tampering with Tradition: The Unrealized Authority of Democratic Agency,* co-edited with Peter Bogason and Sandra Kensen.

Jerry Mitchell is professor of public affairs at Baruch College, The City University of New York. His is the author of a new book published by SUNY Press, *The Business of BIDS*.

Changhwan Mo is currently a research fellow at the Korea Transport Institute and has been advisor at the Regulatory Reform Group in the Prime Minister's Office in South Korea. He is the author or co-author of several articles in the areas of public policy, budgeting, and globalization.

Wayne Parsons is professor of public policy at Queen Mary, University of London. Amongst his publications are *The Political Economy of British Regional Policy; The Power of the Financial Press: Keynes and the Quest for a Moral Science,* and *Public Policy* and he is editor of the New Horizons in Public Policy series for Edward Elgar.

Deike Peters is currently a German Research Foundation (DFG) fellow with the Center for Metropolitan Studies at the Technical University in Berlin. She has a Ph.D. in planning and policy development from Rutgers University and master's degrees in urban planning and international affairs from Columbia University.

Helga Pülzl is currently a post-doctoral researcher at the Department of Economics and Social Sciences at the University of Natural Resources and Applied Life Sciences, Vienna (BOKU). In addition she is a lecturer in comparative politics at the Department of Political Science at the University of Vienna.

Jörg Raab is assistant professor of policy and organisation studies at Tilburg University, the Netherlands. His research focuses mainly on governance mechanisms in the state, economy and society and on different topics in organization theory with an emphasis on inter-organizational networks.

Bernard Reber is research fellow on moral and political philosophy at CNRS-University Paris V. He has also taught at l'Ecole Nationale Supérieure des Mines de Paris, Sorbonne. He is the coeditor of *Pluralisme moral, juridique et politique* and *Les sciences humaines et sociales à l'heure des TIC*.

Donijo Robbins is associate professor for the School of Public & Nonprofit Administration at Grand Valley State University in Grand Rapids, Michigan, where she teaches graduate and undergraduate courses in public budgeting, financial management, and research methods. She holds a Ph.D. in public administration from Rutgers University.

Paul A. Sabatier is professor in the Department of Environment and Policy at the University of California, Davis. He has published *Theories of the Policy Process.*

Alan R. Sadovnik is professor of education, public affairs and administration, and sociology at Rutgers University. Among his publications are *Equity and Excellence in Higher Education; Exploring Education: An Introduction to the Foundations of Education*; and *Knowledge and Pedagogy: The Sociology of Basil Bernstein*.

Thomas Saretzki is professor of environmental policy and politics at the Center for the Study of Democracy, University of Lueneburg (Germany). Currently he is visiting research scholar at Northwestern University in Evanston, Illinois.

Anne Larason Schneider is professor, School of Justice and Department of Political Science, Arizona State University, Tempe. She is co-editor (with Helen Ingram) of *Deserving and Entitled* and co-author (also with Helen Ingram) of *Policy Design for Democracy* (University Press of Kansas, 1997).

Mary Segers is professor of political science at Rutgers University. Her books include *A Wall of Separation? Debating the Role of Religion in American Public Life* (1998) and *Abortion Politics In American States* (1995, co-edited with Timothy Byrnes).

Mara S. Sidney is Associate Professor of Political Science at Rutgers University, Newark. She is the author of *Unfair Housing: How National Policy Shapes Local Action.*

Diane Stone is Marie Curie Chair in the Center for Policy Studies at the Central European University in Budapest, and reader in Politics and International Studies at the University of Warwick. Among her books is *Global Knowledge Networks and International Development* (with Simon Maxwell). She co-edits the journal *Global Governance.*

Eileen Sullivan is a lecturer of political science at Rutgers University. She has been a research director for the New York City Department of Employment, the U.S. Government Accountability Office (GAO), and the Vera Institute of Justice; and she has served as research consultant to the New York City Economic Development Corporation.

Douglas Torgerson is professor of politics at Trent University in Canada. He is a past editor of the journal *Policy Sciences*, and his publications include several critical studies on the theory and history of the field.

Oliver Treib is assistant professor in the Department of Political Science, Institute for Advanced Studies, Vienna. His research topics include EU social policy, new modes of governance and political cleavage structures in international politics.

Michel J.G. van Eeten is an associate professor in the School of Technology, Policy and Management, Delft University of Technology, the Netherlands. He is also a winner of the Raymond Vernon Prize of the Association for Public Policy Analysis and Management and the author (with Emery Roe) of *Ecology, Engineering, and Management.*

Danielle M. Vogenbeck, Ph.D., public affairs, University of Colorado at Denver, is an associate behavioral scientist at RAND, where she specializes in applying social network analysis to organizational change, network governance, and community development projects.

Hendrik Wagenaar is a professor of public policy with the Department of Public Administration at Leiden University. He is the author of *Government Institutions* (Kluwer) and co-editor (with M. A. Hajer) of *Deliberative Policy Analysis* (Cambridge University Press).

Peter Wagner is professor of social and political theory at the European University Institute in Florence, Italy, and professor of sociology at the University of Warwick, UK. His recent book publications include *Varieties of World-Making: Beyond Globalization* (co-edited with Nathalie Karagiannis, 2006) and *A History and Theory of the Social Sciences.*

Christopher M. Weible is an assistant professor in the School of Public Policy, Georgia Institute of Technology in Atlanta. His research interests focus on policy processes and environmental politics, and his work has been published in the *Policy Studies Journal*, *Political Research Quarterly*, and *Journal of Public Administration Research and Theory*.

Kai Wegrich is senior policy analyst at RAND Europe. He received his Ph.D. from Potsdam University. His areas of special interest include public sector reform and regulation.

Hellmut Wollmann is professor (emeritus) of public policy and public administration at the Institute of Social Science of Humboldt University, Berlin, Germany. He was a co-founder and president (1998/1999) of the European Evaluation Society. He is editor of *Evaluation in Public Sector Reform* (2003, with V. Hoffmann-Martinot), *Comparing Public Sector Reform in France and Germany* (2006), and *The Comparative Study of Local Government and Politics* (2006, with H. Baldersheim).

Kaifeng Yang is assistant professor in public administration at Askew School of Public Administration and Policy, Florida State University. He is research associate at the National Center for Public Productivity at Rutgers University and the DeVoe Moore Center for Economic Development at Florida State University.

Dvora Yanow holds the Strategic Chair in Meaning and Method at the Vrije Universiteit, Amsterdam. She is the author of *How Does a Policy Mean?; Conducting Interpretive Policy Analysis; Constructing American "Race" and "Ethnicitiy"* and co-editor of *Knowing in Organizations and Interpretation and Method: Empirical Research Methods and the Interpretive Turn*.

Contents

23 Quantitative Methods for Policy Analysis

Kaifeng Yang

INTRODUCTION

Policy analysis involves using quantitative and/or qualitative techniques to define a policy problem, demonstrate its impacts, and present potential solutions. It often requires sophisticated methods to assess how identified policy problems are impacted by numerous variables, including both policy interventions and contextual factors. Quantitative methods help demonstrate whether a relationship exists between policy designs and policy outcomes, test whether the relationship can be generalized to similar settings, evaluate magnitudes of the effects of policies on social, economic, and political factors, and find better policy alternatives. The use of such methods is part of the scientific expertise with which policy analysts claim their relevance. Techniques such as modeling, quantification of inputs and outputs, descriptive statistics, statistical inference, operations research, cost-benefit analysis, and risk-benefit analysis are frequently used in policy studies.

This chapter discusses the use of quantitative methods in policy analysis. It aims to provide a general understanding of the use of quantitative methods in policy analysis, using examples from the policy analysis literature and linking quantitative methods with the development of policy analysis as a profession and an applied discipline. The chapter has two major sections. The first section briefly reviews the emergence and evolution of quantitative methods in policy analysis, discussing their origin, change, use, and education. The second section introduces some quantitative methods that are widely used in policy analysis.

Due to page limits, this chapter does not cover such basic topics as sampling, level of measurement, reliability, validity, and hypothesis testing, nor does it go into the details of the statistical procedures. Interested readers may find the details in many research methods textbooks. This chapter does not address several important quantitative analysis methods either, such as cost-benefit analysis, survey research, evaluation research, Q methodology, and environment impact assessment, since they are dealt with in other chapters of this handbook. For the same reason, the debate between positivist and post-positivist perspectives is not elaborated here.

HISTORY OF QUANTITATIVE METHODS IN POLICY ANALYSIS

The need for quantitative policy analysis reflects elected officials' desire to design better policies, understand how policies have performed, and assess what impacts policies have made. The use of quantitative methods in policy analysis has its intellectual roots in Harold Lasswell (1951, 1970, 1971), who envisions an overarching policy science discipline based on social science knowledge and methods to analyze policy choices and decision making for the democratization of the society. Policy science, as a multimethod, multidisciplinary, and problem-oriented filed, is concerned with mapping the policy process, policy alternatives, and policy outcomes. Like other social science disciplines, it has to use analytic methods to model policy dynamics and solve policy problems.

Quantitative Analysis as Policy Analysis: 1950s–1960s

Quantitative methods have long been used in public decision making. The New York Bureau of Municipal Research in the 1910s started to use social science methods to systematically study urban problems. In 1922, the Bureau of Agricultural Economics was created within the U.S. Department of Agriculture to examine the relationships between agriculture and the economy and to develop better economic policies. However, more sophisticated use of quantitative methods did not emerge until World War II. The Office of Scientific Research and Development was established in 1941 to coordinate scientific activities during War World II. The Employment Act of 1946 created the Council of Economic Advisors, the first step as Congress formally acknowledged that the executive branch should utilize expert knowledge.

Regarding quantitative methods being used, World War II was a watershed that stimulated new analytic techniques such as systems analysis and operations research. Social scientists began to play more important roles in government decision making by adopting positivism and normative economic reasoning. The economic models dominated the field. For example, scientists and engineers in Great Britain created operations research in World War II in order to help effectively allocate and mange military resources. The technique became widely used in the United States in the early 1950s. It has also been called as management science, systems engineering, and cost-effectiveness analysis. The Rand Corporation, founded in 1948 to do policy analysis work for the government especially the Department of Defense, finally developed the technique into systems analysis, a tool used in the military throughout the 1950s. It was quite successful in solving simple and some complex problems such as inventory management, production scheduling, equipment reliability assessment, and investment risk minimization (Brewer and deLeon 1983).

The 1960s became a "Golden Age" for systems analysis and policy analysis. During this time, policy analysis was essentially quantitative analysis and the research emphasis was on methodology rather than on subject matter. Policy analysis expertise or specializations were in the quantitative methods and techniques, not in their application in specific policy areas. As Radin (2000) observed, professional papers and conferences in the 1960s primarily addressed quantitative analytic procedures such as linear programming, Markov analysis, dynamic programming, game theory, stochastic modeling, Bayesian analysis, quasi-linearization, invariant embedding, and general systems theory. One reason for the quantitative orientation was that most policy analysts during this time were experts in economics. Radin (2000) observed that most of the policy analysts of the time, who were trained as economists or operations researchers, had Ph.D.'s in those areas. Most policy analysis positions were on economic analysis. For example, the Bureau of Agricultural Economics was the center for economic policy research. The Council of Economic Advisers was another prominent policy analysis organ. During this time, policy analysis was influenced by the methodology development of other disciplines, such as the positivism in social science generally, econometrics in economics, and the behavioral revolution in political science.

The domination of quantitative methods in policy analysis during this period was also apparent in the practice. To a large extent, the use of Planning, Programming, and Budgeting Systems (PPBS) is characteristic of policy analysis at this stage. Actively promoted by President John Kennedy's Secretary of Defense, Robert McNamara, PPBS had antecedents in the work of the Rand Corporation. McNamara invited Charles Hitch from Rand to establish a Systems Analysis Unit with responsibility for the PPBS process linking planning with budgeting. The research unit also introduced some other quantitative methods to the federal government such as cost-benefit analysis, operations and systems research, and linear programming. President Johnson, in 1965, required all federal agencies to prepare planning documents and issue-analysis papers to back up their recommendations to the Bureau of Budget. PPBS consisted of three main types of reports: (1) program memoranda, comparing the cost and effectiveness of major alternative programs and describing the agency's

strategy; (2) special analytic studies on current and long-term issues; (3) program and financial plans, summarizing agency outputs, costs, and financing needs over a five-year period. In 1965, the Bureau of the Budget issued a directive to all federal departments and agencies, requiring them to establish central analytic offices that would apply PPBS. In 1969, the National Environmental Policy Act mandated impact analysis in environment policy making.

From the very beginning, statistics has been a curricular requirement. Policy analysis program was thought to help students establish a sense of critical awareness for the general utility of quantitative information (Leinhardt 1981). The early policy literature introduced systems analysis and operations research methods, especially as applied in the defense area (Hitch 1965; Quade 1966; Quade and Boucher 1968). Contents such as how to apply operational research methods, welfare economics, and cost-benefit analysis were common topics in popular textbooks on policy methods during the time.

In public affairs or policy programs, which were first established in the late 1960s, economics was the primary theory, coupled with a number of quantitative techniques. For example, at the University of Minnesota's School of Public Affairs, economics was the core of the required curriculum. Its policy analysis core sequence includes cost effectiveness analysis and PPBS. It also had a quantitative methods sequence teaching the logic of inference and regression analysis (Brandl 1976). In 1968, the University of Michigan reorganized its Institute of Public Administration into the Institute of Public Policy Studies. The program had eight core courses for first year students. Among them, four courses are analytical tools such as statistics, micro and macro economics, cost benefit analysis, and systems analysis. Other course included two in organizational theory and two in political theory and institutions. The purpose was to help students combine latest tools of problem solving and quantitative analysis with a subtle understanding of the social, political, and economic contexts (Walker 1976).

Use of Quantitative Policy Analysis Methods: 1970s–1980s

The use of quantitative techniques such as PPBS had its critics since its emergence. Wildavsky (1969) called for rescuing policy analysis from PPBS, arguing that preconditions for successful PPBS implementation usually do not exist in government. In fact, three years after President Johnson's announcement of a government-wide PPB system, President Nixon issued a memorandum abolishing it. By the 1970s, many limitations of the positivist approach have been acknowledged. In general, those quantitative techniques failed to effectively deal with many complex social problems because those problems cannot be represented with a rational scientific model and do not have a single unitary goal. Operations research places a heavy burden on mathematicians and quantitative analysts to come up with mathematical representation, while overlooking qualitative and soft data, concepts, and methods (Brewer and deLeon 1983). Systems analysis, heavily relying on economics and objective measurements and proxies, does not work well when a full range of human values, interests, and perspectives are considered. Other tools such as flow charts and decision trees were found helpful when there were agreed-upon goals and values. But in reality, policies tend to have multiple and conflicting goals.

Nevertheless, in the 1970s and 1980s, quantitative methods stemming from the systems analysis framework were still widely used and economic models remained dominant although other techniques were also drawn from positivist social sciences. Radin (2000) concluded that "despite the differentiation in practice, the economists' framework, drawn from the market model, continued to dominate" (p. 113). The cost-benefit analysis was extensively used to quantify costs and benefits of policy solutions and thus identify the solution providing the greatest net benefit. For example, the California Legislative Analyst's Office conducted cost-benefit studies for all legislation before the

legislature during the 1970s and 1980s. Cost-benefit analysis was required in the federal government in the 1970s and 1980s for all proposed regulations to be issued by agencies, although the benefits of health, education, and welfare programs are diverse and often intangible. The Executive Order 12291 signed by President Reagan, required detailed cost-benefit analyses for all new federal regulations to assure that federal regulatory agencies thoroughly study the impact of proposed regulations on all concerned parties before promulgation. The order specifies that administrative decisions shall be based on adequate information concerning the need for and consequences of proposed government action, and regulatory objectives shall be chosen to maximize the net benefits to society.

The development of quantitative methods in policy analysis was affected by several historical social events. For example, the Energy Crisis impelled academia to set up energy supply and demand models based on mathematics. The War on Poverty generated a series of new social welfare programs that produced great opportunities for policy analysis. As a result, professional journals and research institutes in public policy were created in a large amount. Significant resources were available for evaluation studies sponsored by the federal government. In the 1950s and 1960s, policy analysis was primarily prospective in that it attempted to assess policy alternatives before a program was actually established. Retrospective policy analysis, which evaluates the impact of an established program, was used in the 1960s but did not become a common practice until the 1970s. At the same time, program failures of some Great Society initiatives prompted policy analysts to address the implementation issues during the policy design stage (Nakamura and Smallwood 1980). Analysis on implementation and program impact entailed the use of more sophisticated methods in order to include more contextual variables. While the measurement of efficiency and filed study were emphasized in the 1960s, experimental studies became important in the 1970s, when social experiments such as Negative Income Tax and Head Start programs were widespread (Daniels and Wirth 1983).

In general, during the two decades, analytic capacity was significantly enhanced due to greater demands for policy analysis, stronger computing capabilities, and advances in economic modeling such as micro-analytic simulation models. However, although policy analysis became more sophisticated, its limits were also exposed (May 1998). The debates between qualitative and quantitative methods, positivist and post-positivist approaches also took momentum. Quantitative techniques were no longer the sole set of skills for policy analysts, and many people realized that political skills were as important as technical skills (i.e., Meltsner 1976).

With support from private foundations, in the late 1970s, public policy graduate programs were set up at Harvard, the University of California at Berkeley, Carnegie-Mellon, the Rand Graduate Institute, the University of Michigan, the University of Pennsylvania, the University of Minnesota, and the University of Texas at Austin. At the University of Michigan, the policy program introduced several new courses on advanced analytical techniques such as modeling and forecasting, policy evaluation, and operations research with emphasis on statistical decision theory. A math refresher course was added to prepare students for advanced statistics (Walker 1976). The National Association of Schools of Public Affairs and Administration established policy analysis as one of five fundamental subject areas. Wildavsky (1976) described the principles for graduate education of public policy based on Berkeley's experiences in the 1970s. He emphasized the importance of multiple analytic perspectives and techniques, arguing that no single set of operations can be taught as the essence of analysis. He viewed analysis as a traveling skill of creatively applying analytic tools to solve various policy problems in a short time period.

Engelbert (1977) reviewed the experiences of policy graduate programs in the early and middle 1970s and pointed out that there was a core subject matter built around: (1) quantitative methodology including mathematical programming and modeling and descriptive and inferential statistics; (2) the political and institutional environment of policy formulation and implementation; (3) economic theory and analysis with emphasis on public-private sector relationships in the allocation of resources; (4) behavioral and nonbehavioral decision making and implementation strategies and

processes; and (5) program management, control, and evaluation. There was a heavy reliance upon quantitative tools of evaluation: "Not only is training in quantitative methodology emphasized in course subject matter . . . but students are expected to demonstrate some proficiency in the application of quantitative techniques to problem-solving exercises" (Engelbert 1977, 231). Intensive instruction was given to techniques such as operations research, model building, cost-benefit analysis, and linear programming.

Democratization for Policy Analysis: 1990s~

In the 1990s, quantitative analysis became far more common and informed, largely because statistical software, such as SPSS, SAS, and STATA, facilitated the use of quantitative methods to deal with complex models and huge datasets. Those statistical packages can calculate numerous statistics and allow their user to manipulate the dataset and transform the variables. Today, policy analysis bears the imprint of the positivist heritage, which is evident in the curricula of policy schools requiring various statistics as core courses. The power of the heritage is also seen in the journals such as *Journal of Policy Analysis and Management* (*JPAM*), *Policy Studies Review* (*PSR*), *Review of Policy Research* (*RPR*), among others. These journals are filled with policy studies using various statistical analyses of particular policies. It is also evident in the annual conferences sponsored by the Association of Public Policy and Management, which, in recent years, have been dominated by papers that employ positivist economic and other research models (Durning 1999).

Meanwhile, there have been methods wars between quantitative and qualitative research; between internal and external validity; and between experimental and statistical control (Brewer 1983; Krane 2001). The quantitative versus qualitative debate reflects the larger battle between "positivists" and "post-positivists." Since the 1980s, the rational positivist approach to policy analysis has been criticized on many grounds. The basis of quantitative methodologies is empirical falsification through objective hypothesis testing of rigorously formulated models. The fundamental positivist principle in policy analysis is to separate facts and values, by which normative issues are translated into technical considerations. In pursuit of replicable relationships, positivists emphasize empirical research designs, causal modeling, scientific sampling, and quantification of outcomes. However, when studying social phenomenon, we can not isolate ourselves from the objects of the research, nor can we separate facts from values.

In the methodology curriculum, positivism equips the students with empirical research designs and statistical methods. Many writers criticize that students trained in this tradition often have little training in understanding the normative and interpretive foundations of the tools they have learned, as well as the social settings to which these techniques are to be applied (Fischer 1998). Therefore, post-positivism has been proposed as an alternative, which is treated as a marriage of scientific knowledge with interpretive and philosophical knowledge about norms and values. In terms of epistemology, post-positivism incorporates deliberative theories and democratic participation.

The tension between positivism and post-positivism has not faded away. On the one hand, one can justifiably argue that positivism is feeble in the face of intractable or wicked problems (Fischer 1995). On the other hand, it is not clear whether post-positivism can specify a common goal of its own and offer their own set of solutions, especially in the operational aspects of policy research (deLeon 1998). Nevertheless, since the 1990s, more efforts have been made to democratize the policy analysis design and process. Participatory design, stakeholder involvement, citizens' input, qualitative methods, and mixed methodology, among others, have contributed to an area with a multidisciplinary theoretical and methodological base (Krane 2001).

Currently, positivism still constitutes the discipline's intellectual infrastructure and is supported by the training, practice, and specialization of the academicians who teach policy analysis methods (Durning 1999). Morçöl (2001) finds that there is considerable support for positivism

among policy professionals, especially among practitioners and professionals with educational background in economics, mathematics, and science. Policy analysis skills in the 1990s include: case study methods, cost-benefit analysis, ethical analysis, evaluation, futures analysis, historical analysis, implementation analysis, interviewing, legal analysis, microeconomics, negotiation and mediation, operations research, organizational analysis, political feasibility analysis, public speaking, small-group facilitation, specific program knowledge, statistics, survey research methods, and systems analysis (Radin 2000).

Vijverberg (1997) recommends that a quantitative methods curriculum should include: (1) Course 1: introduction to probability theory, hypothesis testing, statistical distributions, difference of means test, ANOVA, and rank tests; (2) Course 2: research design and survey methods; (3) Course 3: introduction to regression analysis; (4) Course 4: continuation of regression analysis including maximum likelihood estimation, logit/probit, tobit, simultaneous equations, factor analysis, and LISREL models; (5) Course 5: advanced topics in research methods, including Box-Jenkins (ARIMA), unit roots and cointgration, the introduction to nonparametric statistics, and sample selectivity models. In addition, the economic analysis and operational research traditionally are essential to quantitative policy analysis, so we add them as another category of courses. It can be described as advanced topics in economic analysis and operational research, which includes macroeconomics, microeconomics, cost-benefit analysis, econometrics, operations research, and applied economics.

Take the Master of Public Policy program in the Harris School of Public Policy, University of Chicago as an example, students must finish required and elective courses including Mathematical Preliminaries, Statistical Methods for Policy Research I, Survey Research Methodology, Survey Questionnaire Design, Statistical Methods for Policy Research II, Applied Regression Analysis and some economic analysis courses. Students use computer programs to apply these techniques to real situations (e.g., the effect of sales taxes, labor market discrimination, and redistributive programs). It is also apparent from the curricula and syllabi that economic analysis dominates the teaching for policy analysis.

QUANTITATIVE STATISTICAL METHODS

Statistics is the theory and procedure of analyzing quantitative data obtained from samples of observations in order to study and compare sources of variances of phenomena, to help make decisions to accept or reject hypothesized relationships. Descriptive statistics enable policy analysts to summarize data effectively and meaningfully. Inferential statistics is the use of quantitative techniques to generalize from a sample to a population. In order to choose the right technique policy analysts have to consider the research purpose, the sample size, the distribution of the data, the number of dependent and independent variables, and the type of measurement scale employed by the variables. One can refer to other statistical books for detailed information (i.e., Hair et al. 1998).

Univariate and Bivariate Analysis

Univaraite or descriptive statistics summarize a body of raw data so that the data can be more easily understood. Before descriptive statistics are calculated, policy analysts sometimes use graphs and tables to map the data and have a general sense of the data. For example, frequency polygon displays the trend, Ogive (cumulative frequency polygon) shows percentage of cases following below or above a standard, and both of them can be used to compare different samples. Histograms and bar charts help demonstrate differences among subgroups. Percentages can be calculated to show the proportions, such as the percentage of welfare receipts who are satisfied with the service. Those proportions are sometimes difficult to interpret—as too high or too low, for example—if policy

analysts are not familiar with the context. A 5 percent dissatisfaction rate can be interpreted either as an alarming sign or as prove of quality.

Measures of central tendency indicate the typical value of the data. The mean is the arithmetic average and affected by extreme values. Thus it is not useful for a skewed distribution such as income. The median is the middle observation in a rank-ordered dataset and is insensitive to the observations' values but sensitive to sample size. The mode is the most frequent value, insensitive to the values and sample size. Researchers should find out whether the appearance of two or more modes is due to the mixing-together of different subgroups (e.g., weights of third graders and their parents) in one dataset. The relative value of the mean, median and mode inform policy analysts the shape of the distribution. The choice of an appropriate measure depends on not only the distribution, but also the level of measurement and the analysts' purpose. Also important are measures of dispersion, which sometimes indicate reliability, consistency, and safety. For example, decision makers may be interested in which police department has shorter average emergency response time, but they should also be interested in how consistent those departments are. Analysts should use several descriptive statistics to summarize different aspects of their data to produce a clearer picture. The standard deviation is the most commonly used measure, although it is not useful for a skewed distribution. Another measure, the Inter-quartile range (the distance between the upper and lower quartiles), is hard to calculate mathematically but useful for a skewed distribution.

Bivariate analysis tests whether and how one variable is statistically related with another variable. It helps demonstrate the existence, statistical significance, the direction, and the strength of the relationship. The procedure depends on the level of measurement of the independent and dependent variables. When the independent and dependent variables are categorical (nominal or ordinal), contingency table analysis (cross-tabulation) is generally used. When the independent variable is categorical and the dependent variable is interval or ratio, the difference of means test (t test) or analysis of variance (ANOVA) is preferred. When both variables are interval or ratio, correlation or regression is conducted.

In contingency table analysis, analysts first separate the observations into groups based on their values for the independent variable, then calculate percentages within the independent categories, and finally compare the percentages across one of the dependent categories. The percentage difference tells analysts whether the independent variable makes a difference (Meier and Brudney 2002). The chi-square (χ^2) test is then used to assess the statistical significance of the relationship—whether we can reject the null hypothesis that assumes no relationship between two variables in the population based on our sample observations. Chi-square test indicates the probability that the results can be generalized to the population. However, chi-square is not a measure of substantive importance or strength because chi-square result is affected by the sample size: if the sample size N is large (say, greater than 1, 500), χ^2 will usually be large even if the association is weak. The importance or the strength of the relationship is better measured, especially when dealing with large samples, by a measure of association that ranges from +1.0 (prefect positive relationship) and –1.0 (perfect negative relationship). When both variables are ordinal, the most frequently used measures of association are Kendall's tau-b (for square tables), Kendall's tau-c (for non-square tables), Somer's d, and Goodman and Kruskal's gamma. In general, the tau measures are used more commonly then the Somer's d measures. Many analysts often use both gamma and either tau-b or tau-c. When one or both of the variables are nominal, Goodman and Kruskal's lambda should be used.

The difference of means test and the analysis of variance have similar logic. Analysts first divide observations into categories based on the values of the independent variable. A relationship exists if the values of the dependent variable are quite different across groups and have smaller within-group variance than before (Johnson and Reynolds 2005). To determine the statistical significance, the difference of means test uses t test and compares the result with the appropriate criterion (large t values lead to rejection of the null hypothesis). The analysis of variance uses F statistic to measure the statistical significance. F is the ratio of between-group mean square to

within-group mean square. The *F* ratio is compared with an *F*-ratio table to decide whether to reject the null hypothesis.

Linear regression, or ordinary least squares regression, is to find the best line function to describe a relationship that can minimize the squared errors. Its general form is $Y = a + bX + e$, in which a is the intercept, b is the slope, and e is the error term. The formula for b, the regression coefficient, is

$$\frac{\sum(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sum\,(x_i - \bar{x})^2.}$$

The coefficient shows how much the estimated average Y value will change if X is changed one unit. The goodness of fit may be measured by the standard error of the estimate, which indicates the amount of error that one makes when predicting a Y value for an X value. Another common goodness of fit measure is the coefficient of determination (r^2) ranging from zero (lack of fit at all) to one (perfect fit). The coefficient of determination is the ratio of the explained variation to the total variation in Y, or the ratio of the reduction of the error by using the regression line to the total error by guessing the mean.

Analysis of Variance (ANOVA)

ANOVA is a dependence technique that explains the variation of a metric dependent variable based on a set of nonmetric (categorical) independent variables. Its general form is:

$$\underset{\text{(metric)}}{Y_1} = \underset{\text{(nonmetric)}}{X_1 + X_2 + X_3 + \ldots + X_n}$$

ANOVA helps determine whether samples from two or more groups come from populations with equal means. It is a primary tool for analyzing experimental data. ANOVA examines within-groups variances (MS_W) and between-groups variances (MS_B). The ratio of MS_B to MS_W, which is the *F* statistic, measures how much variance is attributed to the different treatments versus the random sampling error. Large values of the *F* statistic lead to rejection of the null hypothesis assuming no treatment effects. Consider a study on job training and job placement. Policy analysts want to compare the effect of two programs, School Training and On-the-Job Training (OJT), on job placement in terms of salary. Participants are randomly assigned to one of the two programs. ANOVA is used to test the difference between the two groups. The independent variable here is the type of job training programs and the dependent variable is participants' monthly earnings. ANOVA is widely used in policy studies. For example, Wells, Layne, and Allen (1991) used ANOVA to assess whether learning styles differed for supervisory, middle, upper middle, upper, and executive managers in the Georgia Department of Corrections.

To use ANOVA, the data has to satisfy the assumptions of linearity, normality (the dependent variable is normally distributed), and equal variance (variances are equal for all treatment groups). However, *F* tests in ANOVA are robust regarding these assumptions except in extreme cases. The equal variance assumption is often ignored if the number of cases in each group is similar. Analysts are encouraged to examine the data first to assess the presence of nonlinear relationships and outliers.

Multiple Regression Analysis

Multiple regression analysis examines the relationship between a single metric dependent (criterion) variable and a set of metric independent (predictor) variables. Its general form is:

$$\underset{\text{(metric)}}{Y} = \underset{\text{(metric)}}{\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n + \varepsilon}$$

α is a regression constant, representing the value of Y when all the independent variables have values of zero. β is a regression coefficient indicating the relationship between X and Y controlled for all other independent variables. ε is an error term that incorporates the cumulative effect on Y of factors not included in the model. Regression may be used to calculate the predicted value of Y for any given value of X. And the residuals or distances between the predicted and observed values of Y lead to a measure (R^2) of how well the equation fit the data.

Regression analysis is the most widely used and versatile dependence technique in policy analysis for the purpose of prediction or explanation. For example, regression analysis is the foundation for forecasting models that predict national economy or other performance based on certain inputs. It is used to examine how decisions are made and how attitudes are formed. It is also used to identify quality determinants of policy implementation and program design. Hunter (2001) used multiple regression to explain the difference of states' economic growth by lobbying efforts in selected categories and a sample of demand-side economic policies, controlling for net business growth, expenditure, and republican control of the government and legislature. The economic growth was measured by the change of a state's per capita gross state product (GSP) between 1986 and 1991. The regression results suggest that two categories of lobbying efforts explain more of the variance in GSP than the demand-side policies and the other variables. With multiple regression, the author was able to show that the control variables contributed to 9 percent of the variation in changes in GSP while the dependent variables contributed to an additional 34 percent variation.

A very important but often ignored step is to assess whether the model satisfies the assumptions of regression analysis, such as existence, linearity, homoscedasticity (equal residual variances), independence of the residuals, and normality. The principal diagnostic method is to examine the residual—the difference between the actual dependent variable value and its predicted value—though partial regression plots and statistical tests (i.e., the Kolmogorov-Smirnov test and the Shapiro-Wilks test for normality; the Durbin-Watson test for independence). In graphical analysis, a triangle-shaped or a diamond-shaped pattern indicates the presence of heteroscedasticity, which can also be assessed with the Levene test in SPSS. Also important is to avoid multicollinearity, which can substantively affect explanation and estimation of the regression coefficients and their significance tests. Analysts can use correlation matrix for the independent variables to observe whether high correlations are present (.90 and above). More common measures are the tolerance value and the variance inflation factor (VIF, rule-of-thumb cutting value at 10.0), which measures the degree to which each independent variable is explained by the other independent variables. Analysts may use some remedial strategies to solve the above problems, and data transformation (i.e., from Y to log Y or Y^2) is one of the options. In the final steps, analysts also need to identify outliers and determine whether they should be excluded from the analysis. Common indicators for this purpose are the leverage *h* and the Cook's distance, which measures the extent to which the regression coefficients change when the particular observation is deleted.

Time Series Analysis

Time series analysis identifies the pattern of change across time in order to explain the phenomenon and to predict the future based on historical and existing patterns. It enables policy analysts to examine a variable, such as unemployment rate and economic growth, over equally spaced intervals of time such as month and year. Its general form is

$$X_t = a_1X_{t-1} + a_2X_{t-2} + \ldots + a_nX_{t-n}$$

Time series analysis is important to policy analysis since policy change is a crucial question and time series analysis permits data-based forecasting. Many policy studies are cross-sectional, and the results may be strengthened by replicating the study in different times. In addition, time-series analysis can address questions of causation that would be impossible to tackle with cross-sectional

analysis, given that the temporal sequencing of changes can be established with a time-series. For example, to answer the question whether the incidence of crimes in a region changed following the establishment of a new crime-fighting program, interrupted time-series experiment is an appropriate strategy. To predict the pattern over time based on Gallup polls of presidential popularity, time-series regression is an appropriate strategy.

In general, time series analysis can serve three purposes: analyses of trends and forecasting; causal analyses; and program and policy analyses (Burbridge 1999). Especially, interrupted time-series is useful because the introduction of a program or policy will produce a break in the time-series trend for certain variables affected by the program or policy. Analysts need to have enough pre-program data to establish a pre-program trend, to know the exact time of the introduction of the program and reasonable assumption about how long it will take for the program to affect the long term trend (Burbridge 1999).

There are six basic steps in a time series analysis. First, plot the data. Second, examine the plot and determine if any short-term fluctuations exist. Third, if the data show a cyclical trend, determine the length of the short-term trend and filter the trend. Fourth, determine whether a relationship exists. Fifth, use linear regression to estimate the relationship between time and the variable being analyzed. Sixth, make a forecast by using the regression equation (Meier and Brudney 2002).

For example, in a research on policy design, bureaucratic incentives, and policy enforcement, Keiser and Meier (1996) hypothesized that local-level implementation environment and resources committed to implementation affect the actual enforcement levels. Using pooled time series data of federal laws on child support enforcement from 1983 to 1991, they were able to confirm the hypotheses. Albritton (1979) measured impacts of the Title XX amendments to the Social Security Act with an interrupted time-series analysis. The Auto-Regressive Integrated Moving Averages (ARIMA) model was adopted and the results showed that the policy innovation led to dramatic, nonincremental changes. Morgan and Pelissero (1980) used an interrupted time-series quasi-experiment to test the hypothesis that reformed cities tax and spend less than unreformed cities. Eleven cities, with a population of 25,000 and above, which reformed their political structure between 1948 and 1973, were compared with eleven matched cities that did not reform. The results showed that government structure did not affect cities' fiscal behavior.

Event History Analysis (EHA)

Event history analysis is used to explain why certain units of analysis (individuals, organizations, or states, etc.) are more likely to experience the event(s) of interest than others. It is a specialized subfield of time series analysis that analyzes rare events (time series in which most data are non-events). The data in EHA measures the number, timing, and sequence of changes in a variable of interest. EHA can be a form of panel study in which the periods of observation are not arbitrarily spaced but instead measurement is taken at each stage of a sequence of events. The dependent variable is qualitative and taking values between zero and one, but the independent variables can take any real numbers.

The key concepts of EHA include a risk set (a set of unit of analysis that have yet to experience a particular event), a survivor function (the decline in the size of risk over time), and the hazard rate (the rate at which particular events occurring at a particular time). EHA assumes that it is possible to predict the dependent variable (e.g., marriage, employment changes, higher education, and death) within certain time frames. The rationale stems from the life table analysis used by demographers to calculate survival and mortality rates in a given population over time. For example, if x number of the population is alive at time t, it is possible to predict the survival rate of that population at time t + 1. The hazard rate in EHA is the other side of the survival rate and refers to the probability of a dependent variable occurring to an individual within a specified time frame, given that individual is at

risk (Cohen, Manion, and Morrison 2000). The problem is solved by taking a logit transformation of the dependent variable and then estimating with maximum likelihood techniques (Allison 1984).

EHA began to be used in social sciences in the 1970s. It was prominent in the field of international relations, where it was used to analyze time series of international conflict and diplomatic events. Policy analysts applied EHA in other areas later on. Plotnick (1983) used EHA to study the entry to and exit from the Aid to Families with Dependent Children (AFDC) program. The estimates were applied to projected changes in lengths of time spent on and off AFDC and in AFDC caseloads due to changes in the dependent variables. The results demonstrated that age and wage have significant, negative effects on the rate of entering AFDC, and significant, positive effects on the exit rate.

Berry and Berry (1990), examining state lottery adoptions, used EHA to explain how states' internal characteristics (political and economic) and regional diffusion influenced the probability that the state adopted a lottery. An EHA model was developed as:

$$ADOPT_{i,t} = \Phi\left(\begin{array}{c} b_1 FISCAL_{i,t-1} + b_2\, PARTY_{i,t} + b_3 ELECT1_{i,t} + b_4 ELECT2_{i,t} + b_5\, INCOME_{i,t-1} \\ + b_6 RELIGION_{i,t-1} + b_7 NEIGHBORS_{i,t} \end{array}\right)$$

$ADOPT_{i,t}$	=	the probability that state i will adopt a lottery in year t
$FISCAL_{i,t-1}$	=	the fiscal health of a state's government in the previous year
$PARTY_{i,t}$	=	the degree to which a political party controls the government
$ELECT1_{i,t}$	=	dummy, 1 for the year of gubernatorial election
$ELECT2_{i,t}$	=	dummy, 1 for neither the year of an election nor the year after
$INCOME_{i,t-1}$	=	personal income
$RELIGION_{i,t-1}$	=	the proportion population adhering to fundamentalism religion
$NEIGHBORS_{i,t}$	=	the number of previously adopting neighboring states

The results showed that previous adoption by neighboring states and declining fiscal health affect the probability of adopting the lottery. The authors noted that lottery adoption was most likely to occur in the years immediately following the election. In addition, states with lower per capita income and states with higher percentage of religious fundamentalists were least likely to adopt lotteries. With EHA, Berry and Berry (1990) concluded that regional diffusion and internal determinants were valid explanations of state lottery adoption. They proposed that EHA should be used in other subfields of political science because it takes advantage of both temporal and cross-sectional variation in political behavior, and it remains valid for rarely occurred events such as wars and switching political party identification. Box-Steffensmeier and Jones (1997) illustrated EHA methods with three issues: overt military interventions, challenger deterrence, and congressional career paths. They called for greater use of EHA models as well.

Factor Analysis

Factor analysis is an interdependence technique in which all variables are simultaneously considered and factors are created to explain the variable set. Factor analysis has three basic purposes: to identify factor structure underlying the variables, to achieve data reduction, and to test the relationships among variables. Factor analysis is based on the fundamental assumption that some underlying factors, which are smaller in number than the number of observed variables, are responsible for the covariation among observed variables. The emphasis on an underlying factor structure reflects a belief that there are real qualities in the world, such as trust, motivation and satisfaction, which

are not directly measurable but can be revealed through the covariation of related variables. Its general form is:

$$X_1 = b_1(F_1) + b_2(F_2) + \ldots + b_n(F_n) + d_1(U_1)$$

where

X_1	=	the subject's score on observed variable 1
b_n	=	the weight for underlying common factor n, as used in determining the subject's score on X_1
F_n	=	the subject's score on underlying factor n
d_1	=	the regression weight for the unique factor associated with X_1
U_1	=	the unique factor associated with X_1

Factor analysis has two types: exploratory and confirmatory. Confirmatory factor analysis is used with path analysis for structural equation modeling. For exploratory factor analysis, if cases are being grouped then it becomes Q method or cluster analysis; if variables are being grouped then it is the R-type factor analysis. Factor analysis differs from the principal components analysis in that the components of principal component analysis account for total variance in the data while the factors of factor analysis account for common variance in the dataset. Factor analysis assumes that the observed variables are linear combinations of the underlying factors. In contrast, principal component analysis assumes that components are linear combinations of observed variables. Therefore, factor analysis can be used to identify the number and nature of the factors that are responsible for covariation in the dataset, but principal component analysis cannot. Nevertheless, many writers do not make the distinction especially when the purpose is to reduce items or variables.

For example, Winter and May (2001) measured Danish farmers' social motivation to comply with regulations with six survey items about farmers' perceptions of the enforcement style of municipal inspectors. They then used the principal component analysis, treated as factor analysis, and identified two underlying dimensions of enforcement style: formalism and coercion. Warner and Hebdon (2001) studied factors affecting local governments' restructuring choices among privatization and its alternatives. In addition to fiscal stress and control variables such as per capita income, municipal type, size of government and tenure of office, the authors developed fourteen items to measure economic and political conditions of the local governments. They conducted principal components analysis and reduced the fourteen items to three distinct components: information and service quality; efficiency; and unionization and political factors. In Table 23.1, the first seven items have factor loadings higher than 0.5 on Information and Service Quality, with lower loadings for the other two components. Therefore, the seven items can be used together in the future analysis. The eighth item, local employment impact, has similar loading on the first component (0.476) and the third (0.452). Therefore, this item should have been deleted from future analysis.

Path Analysis

Path analysis is used to test the indirect and casual relationships among the variables specified in the model. Policy analysts first draw a path diagram based on a theory or a set of hypotheses, then estimate path coefficients using regression techniques, and finally determine indirect effects (Nachmias and Nachmias 1996). It is very useful when dealing with mediating effects, where an independent variable had an impact on an intervening variable which, in turn, had an impact on a dependent variable. Path analysis assumes perfect reliability of the instruments used to operationalize variables. Therefore, all variables in the path model are considered to be observed, not latent or underlying factors. When it is used mathematically with confirmatory factor analysis (CFA), it becomes structural equation modeling (SEM) and can deal with latent variables. SEM allows

TABLE 23.1
Principal Components Analysis Results from Warner and Hebdon (2001)

	Information & Service Quality	Efficiency	Union
Information (1)	0.792	0.17	0.038
Legal	0.643	–0.048	0.407
Community Values (2)	0.614	0.2	0.27
Monitoring (3)	0.613	0.189	0.301
Service Quality (4)	0.604	0.481	–0.003
Leadership	0.563	0.434	–0.009
Experience	0.529	0.125	0.132
Local Employment Impact	0.476	0.196	0.452
Economic Efficiency	0.147	0.832	0.092
Budgetary Impact	0.07	0.793	0.339
Management	0.321	0.693	0.112
Labor	0.457	0.471	0.419
Union	0.076	0.075	0.799
Political (5)	0.216	0.243	0.573

Note: N = 201; Based on a 1997 survey on New York State towns and counties.

assessment of the reliabilities of the latent variables, more precise estimation of the indirect effects of the exogenous variables, and multiple dependent variables.

Path analysis is used to both simplify and depict complex theoretical relationships. LISREL (Linear Structural Relations) has been the popular program since 1981, and statistical packages such as SAS and Stata can conduct the analysis as well. Ellickson (1992) used path analysis to explain the impact of personal, environmental, and institutional factors on legislative success with data drawn from the 1987–88 Missouri House of Representatives. The results showed that institutional variables, seniority and political party, have the strongest impact. The path analysis was able to show that formal office is an intervening variable between legislative success and other independent variables such as age, urbanism, seniority, and political party.

Cohen and Vigoda (1998) used path analysis to compare two different models explaining the relationship between citizenship behavior and work outcomes. Figure 23.1, the direct model, has no mediating variables. The results show that political participation, community involvement, and general altruism have statistically significant direct impact on perceived performance, while disillusionment with government has significant direct impact on turnover intentions. Figure 23.2, the indirect model, has four independent variables (political participation, community involvement, general altruism, and disillusionment with government), one mediating variable (participation in decision making), and two dependent variables (turnover intentions and perceived performance). Among the independent variables, only community involvement has a statistically significant path to participation in decision making. In comparison, model fit indices suggested that the direct model is better than the indirect model.

Game Theory

Game theory is a mathematical approach to individual decision making that employs games as paradigms of rational decision-maker interactions. A game is any interaction between agents that is governed by a set of rules specifying the possible moves for each participant and a set of outcomes for each possible combination of moves. A game of "pure strategy" consists of the following interrelated components: The *players*, who may be people or organizations, choose from a list of *options*

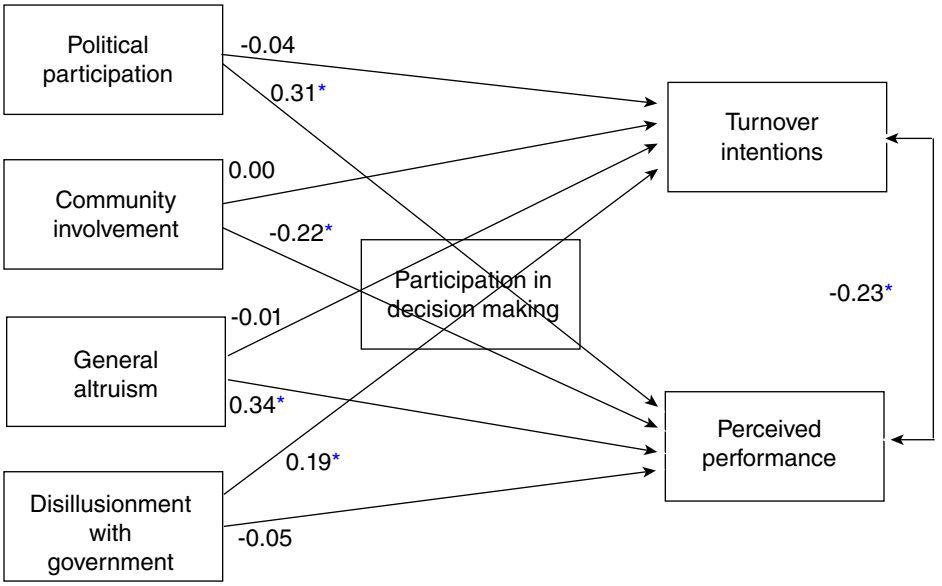

FIGURE 23.1 An Indirect Path Model from Cohen and Vigoda (1998).

or *strategies* available to them**.** At each stage of the play, the players choose their course of *action* from a set of possible decisions, which are not usually the same for each player. The actions lead to *outcomes* or *consequences*. It assumes the players have fixed *preferences* for the outcomes: they like some outcomes more than others. After the decisions have been made, each player receives a certain *payoff* measured in a common unit for all players (Morrow 1994).

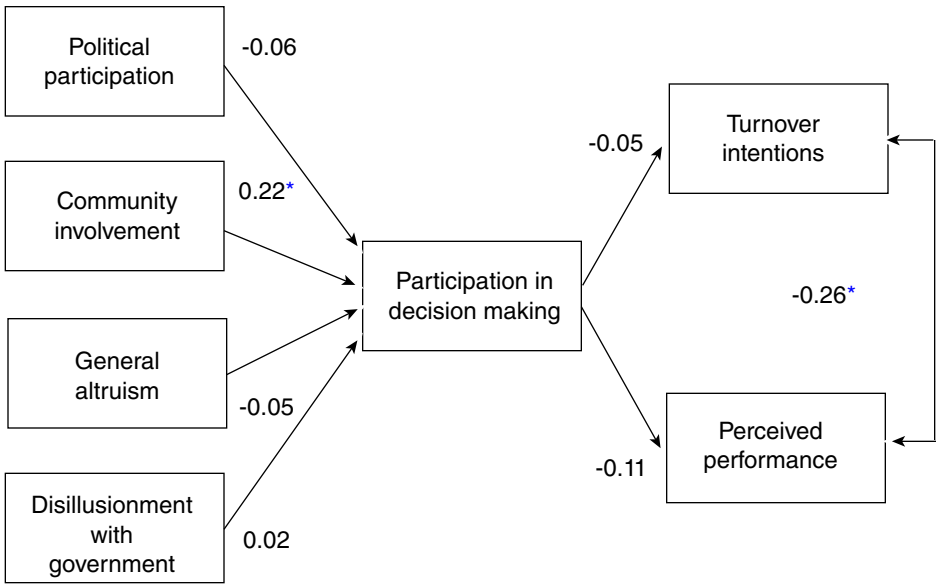

FIGURE 23.2 A Direct Path Model from Cohen and Vigoda (1998).

The assumptions of game theory are: (1) individual action is instrumentally rational, (2) common knowledge of rationality held by all the players, (3) the players will draw the same inferences on how a game is to be played, (4) players know the rules of the game and their motive is independent of the rules, (5) fixed preferences, (6) transitivity (if A>B and B>C then A>C) (Heap and Varoufakis 2004; Gates and Humes 2000). Apparently, those assumptions are simplistic and subject to criticisms. For example, individual identities and preferences may not be pre-fixed; rather, they are socially embedded and constituted. They are often generated during the specific social interactions.

In a policy situation, we may encounter different occurring events that result from decisions made by others. When actors seek to maximize their own interests but their actions affect one another, a game condition involving both conflict and cooperation exists. Game theoretic models help actors make decisions when confronted with competing policy alternatives or decision consequences. Both politics and games involve the moves and interactions of players attempting to maximize their interests; the selection of strategies with specific consequences; and, at times, coalition formation (Kelly 2003).

There are several game forms. The simplest one is the two-person, zero-sum game in which two players are involved and one player's gains are the other player's losses, and vice versa. Consider the Prisoner's Dilemma, one of the classic games, as an example. The two players are partners in a crime who have been captured by the police. Each suspect is placed in a separate cell and offered the opportunity to confess. Each prisoner has two choices: strategy A (confess) or strategy B (do not confess). The payoff for a prisoner in any particular round depends on both prisoners' choices in that round. As shown in the tradeoff table (Table 23.2), there are four possible scenarios: (1) both choose to confess (strategy A), and each of them earns the same payoff of 3; (2) both choose not to confess (strategy B), and each of them has the same payoff of 2; (3) Prisoner 1 chooses to confess (strategy A) while Prisoner 2 chooses not to (strategy B). As a result, Prisoner 1 earns a payoff of 5 while Prisoner 2 earns a payoff of 1; (4) Prisoner 1 chooses not to confess (strategy B) while Prisoner 2 chooses to confess. As a result, Prisoner 1 earns a payoff of 1 while Prisoner 2 earns a payoff of 5.

The Prisoner's Dilemma relates to the issue of trust, the free rider problem, public goods, negotiation, regulation, corruption, and conflict resolution. Axelrod (1984) demonstrated that *Tit-for-Tat,* a program starting with a co-operative move and then following whatever the opponent did on the previous move, is the best strategy in repeated Prisoner's Dilemma games. It indicates that although cooperation is not a Nash equilibrium in the one-shot game, it is in indefinitely repeated games. Both Axelrod's analysis (1984) and Smith's (1982) analysis have led to many other applications in the filed of political science (see Axelrod and Dion 1988). Game theory has been used in political science since the 1950s, especially in areas such as voting, group preference, coalition formation, bargaining, diplomacy, and negotiation (Shubik 1982). After Harsanyi (1967) introduced the concept of incomplete information to game theory in the late 1960s, incomplete information models have been applied to voting, political activism, bureaucratic control, crisis bargaining, arms control agreements, and alliance formation (Gates and Humes 2000).

TABLE 23.2
The Prisoner's Dilemma

		Prisoner 1	
		Strategy A	**Strategy B**
Prisoner 2	Strategy A	(3,3)	(1,5)
	Strategy B	(5,1)	(2,2)

Simulation

Simulation is a quantitative procedure by which analysts build mathematical models of policy process that are difficult to solve analytically and then run the models on a series of organized trial-and-error experiments in order to simulate the behavior of the system over time. It helps analysts understand the system by simulating it in the environment and determining the likely course of events and conditional changes in public policy. It helps answer questions such as: "What would happen to our local economic development policies if the inflation rate is 4 percent instead of 3 percent in the coming year?" Or, "How would this growth management strategy influence the traffic of this county in twenty years?" Simulation sometimes is the only method available if the actual environment or system is difficult to observe or model, or if the model is too complicated to be solved analytically. In some other times, it is infeasible (i.e., too expensive or disruptive) to actually operate and observe a system. For example, if analysts are comparing two ways of providing benefits to veterans, operating two different systems may cause great confusion and legal problems.

A good simulation should satisfy the following conditions: (1) Calibrated. Accurate data are included in the construction of the simulation, and the values for the parameters match empirical observation as closely as possible; (2) Checked. The functioning of the model is comparable to the actual functioning of the real world; (3) Flexible. The model is flexible enough to answer a variety of questions (Kane 1999). The general steps are: (1) define the system one intends to simulate, (2) formulate the model one intends to use, (3) identify and collect data necessary to test the model, (4) test the model and compare its behavior with the actual environment, (5) run the simulation, (6) analyze the results and revise the solution if desired, (7) rerun the simulation to test the new solution, (8) validate the simulation (Levin et al. 1989).

Despite the criticism that it lacks mathematical elegance and precision, simulation is one of the most widely used operations research techniques. In the 1960s, it was used in international relations and urban affair issues such as municipal budgeting, election, and political recruitment (see Coplin 1968). Its use has grown considerably with the development of mathematical modeling and informational technology. It is especially useful in answering "what if…" questions (Zagonel et al. 2004). At the University of Rhode Island, the Department of Environmental and Natural Resource Economics created a Policy Simulation Laboratory (SimLab) to apply interactive tools based on modern computer technologies to help understand the consequences of policy actions. For example, the town council in one of the Group Decision Rooms of SimLab might design a plan for managing growth in the town. Computer systems then simulate the environment and predict the economic and social consequences of the plan. Geographic Information Systems are used to present the consequences for the town with electronic maps.

Simulation is used in a variety of policy settings such as the construction of electoral districts (Gelman and King 1994), the making of foreign policy (Taber 1992), the effects of emission controls on the earth's climate (Bankes and Lempert 1996), social security reform (Weller 2000), and alternate approaches to health insurance expansion (Remler, Zivin, and Glied 2004). Tengs et al. (2004) created a Tobacco Policy Model to examine the potential consequences of mandating tobacco companies to improve the safety of cigarettes. Through simulation of the U.S. population over a fifty-year time span, the results show that even if the safety mandate makes smoking more attractive and increases tobacco use, it is still possible to obtain a net gain in population health. Robins, Michalopoulos, and Pan (2001) used a simulation model to examine whether welfare recipients would work full-time if offered an earnings supplement conditioned on full-time employment. The simulation model extended a traditional microeconomic model of the income or leisure choice to include the choice to receive welfare, assuming that welfare recipients' decisions about employment and welfare were based on the intention to maximize their economic well-being. Outcomes were simulated with three different financial incentives: AFDC (pre-TANF environment), TANF (currently used in the sample states as required by the Temporary Aid to Needy Families Act), and SSP (Self-Sufficiency Project).

The results suggested that the earning supplement would increase full-time employment while the TANF incentives would encourage primarily part-time employment.

CONCLUSION

Quantitative methods help assess the relative and joint effects of a variety of independent variables on some dependent variables. They inform citizens and clients about policy choices with numbers, graphs, and tested relationships. They enable citizens and clients to see the benefit and risks of policy alternatives with mathematical eyes. Development of more sophisticated quantitative techniques is a crucial task for many current policy analysts (Wagle 2000). As policy problems become more complex, environments become more turbulent, and time and budgets become more constrained, policy analysts must be able to choose the most appropriate (valid, reasonable, and realistic) strategy and implement the study in a short period of time.

REFERENCES

Albritton, R. B. (1979). Measuring public policy: Impacts of the supplemental security income program. *American Journal of Political Science*, *23*(3), 559–578.

Allison, P. D. (1984). *Event history analysis*. Newbury Park: Sage.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Axelrod, R., and Dion, D. (1988). The further evolution of cooperation. *Science 242*. 1385–1390.

Bankes, S., and Lempert, R. J. (1996). Adaptive strategies for abating climate change. In Fogel, L. J., Angeline, P. J., & Back, T. (eds.), *Proceedings of the fifth annual conference on evolutionary programming*, pp. 17–25. Cambridge, MA: MIT Press.

Berry, F. S., and Berry, W. D. (1990). State lottery adoptions as policy innovations: An event history analysis. *American Political Science Review*, *84*(2), 395–415.

Box-Steffensmeier, J. M., and Jones, B. S. (1997). Time is of the essence: Event history models in political science. *American Journal of Political Science*, *41*(4), 1414–1461.

Brandl, J. E. (1976). Public service education in the 1970s. *Urban Analysis*, *3*, 105–114.

Brewer, G. D., and deLeon, P. (1983). *The foundations of policy analysis*. Homewood, IL: The Dorsey Press.

Brewer, M. B. (1983). Evaluation: Past and present. In E. L. Struening and M. B. Brewer (eds.), *Handbook of evaluation research*, Beverly Hills, CA: Sage.

Burbridge, L. (1999). Cross-sectional, longitudinal, and time-series data: Uses and limitations. In G. J. Miller and M. L. Whicker (eds.), *Handbook of research methods in public administration*, pp. 283–300. New York: Marcel Dekker.

Cohen, A., and Vigoda, E. (1998). An empirical assessment of the relationship between general citizenship and work outcomes. *Public Administration Quarterly*, *21*(4), 401–431.

Cohen, L., Manion, L., and Morrison, K. (2000). *Research methods in education*, 5th ed., New York: Routledge.

Coplin, W. D. (1968). *Simulation in the study of politics*. Chicago: Markham Publishing.

Daniels, M. S., and Wirth, C. J. (1983). Paradigms of evaluation research: The development of an important policy-making component. *American Review of Public Administration*, *17(1),* 33–45.

deLeon, P. (1998). Introduction: The evidentiary base for policy analysis: Empiricist versus postpositivist positions. *Policy Studies Journal*, 26(1), 109–113.

Durning, D. (1999).The transition from traditional to postpositivist policy analysis: A role for Q-methodology. *Journal of Policy Analysis and Management*, 18(3), 389–410.

Ellickson, M. C. (1992). Pathways to legislative success: A path analytic study of the Missouri house of representatives. *Legislative Studies Quarterly*, 17(2), 285–302.

Engelbert, E. A. (1977). University education of public policy analysis. *Public Administration Review*, 37(3), 228–236.

Fischer, F. (1995). *Evaluating public policy*. Chicago: Nelson-Hall.

Fischer, F. (1998). Beyond empiricism: Policy inquiry in postpositivist perspective. *Policy Studies Journal*, 26(1), 129–146.
Gates, S., and Humes, B. D. (2000). *Games, information, and politics*. Ann Arbor: The University of Michigan Press.
Gelman, A., and King, G. (1994). Enhancing democracy through legislative redistricting. *American Political Science Review,* 88(3), 541–559.
Gottman, J. M. (1988). *Time-series analysis*. New York: Cambridge University Press.
Hair, J. F., Tatham, R. L., Anderson, R.E., and Black, W. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
Harsanyi, J. (1967). Games of incomplete information played by Bayesian players. *Management Science,* 14, 159–182, 320–334, 486–502.
Heap, S. P. H., and Varoufakis, Y. (2004). *Game theory*. New York: Routledge.
Hitch, C. J. (1965). *Decision-making for defense*. Berkeley, CA: University of California Press.
Hunter, K. G.. (2001). An analysis of the effect of lobbying efforts and demand-side economic development policies on state economic health. *Public Administration Quarterly,* 25(1), 49–78.
Johnson, J. B., and Reynolds, H. T. (2005). *Political Science Research Methods* (5th ed.). Washington, D.C.: CQ Press.
Kane, D. (1999). Computer simulation. In G. J. Miller and M. L. Whicker (eds.), *Handbook of research methods in public administration*, pp. 511–533. New York: Marcel Dekker.
Keiser, L. R., and Meier, K. J. (1996). Policy design, bureaucratic incentives, and public management: The case of child support enforcement. *Journal of Public Administration Research and Theory,* 6(3), 337–364.
Kelly, M. A. (2003). Game theory. In J. Rabin (ed.), *Encyclopedia of public administration and public policy*, pp. 533–536. New York: Marcel Dekker.
Krane, D. (2001). Disorderly progress on the frontiers of policy evaluation. *International Journal of Public Administration*, 24(1), 95–123.
Lasswell, H. D. (1951). The policy orientation. In D. Lerner and H. D. Lasswell (eds.), *The policy sciences*, pp. 3–15. Stanford: Stanford University Press.
Lasswell, H. D. (1970). The emerging conception of the policy sciences. *Policy Sciences*, 1, 3–14.
Lasswell, H. D. (1971). *A pre-view of policy sciences*. New York: Elsevier.
Leinhardt, S. (1981). Data analysis and statistics education in public policy programs. In J. P. Crecine (ed.), *Research in public policy analysis and management*, pp. 53–61. Greenwich, CT: JAI Press.
Levin, R. I., Rubin, D. S., Stinson, J. P., and Gardner, E. S. (1989). *Quantitative approaches to management*. New York: McGraw-Hill.
May, P. J. (1998). Policy analysis: Past, present, and future. *International Journal of Public Administration*, 21(6-8), 1089–1114.
Meier, K. J., and Brudney, J.L. (2002). *Applied statistics for public administration* (5th ed.). Belmont, CA: Wadsworth.
Meltsner, A. J. (1976). *Policy analysis in the bureaucracy*. Berkeley: University of California Press.
Morçöl, G. (2001). Positivist beliefs among policy professionals: An empirical investigation. *Policy Sciences*, 34, 381–401.
Morgan, D. R., and Pelissero, J. P. (1980). Urban policy: Does political structure matter? *American Political Science Review*, 74(4), 999–1006.
Morrow, J. D. (1994). *Game theory for political scientists*. Princeton, NJ: Princeton University Press.
Nachmias, C. F., and Nachmias, D. (1996). *Research Methods in the Social Sciences* (5th ed.). New York: St. Martin's Press.
Nakamura, R. T., and Smallwood, F. (1980). *The politics of policy omplementation*. New York: St. Martin's Press.
Quade, E. S. (Ed.). (1966). *Analysis for military decisions*. Chicago: Rand McNally.
Quade, E. S., and Boucher, W. I. (Eds.). (1968). *Systems analysis and policy planning: Applications for defense.* New York: American Elsevier.
Plotnick, R. (1983). Turnover in the AFDC population: An event history analysis. *The Journal of Human Resources*, 18(1), 65–81.
Radin, Beryl A. (2000). *Beyond Machiavelli*: *Policy analysis comes of age*. Washington DC: Georgetown University Press.

Remler, D. K., Zivin, J. G., and Glied, S. A. (2004). Modeling health insurance expansions: Effects of alternate approaches. *Journal of Policy Analysis and Management*, 23(2), 291–313.

Robins, P. K., Michalopoulos, C., and Pan, E. (2001). Financial incentives and welfare reform in the United States. *Journal of Policy Analysis and Management*, 20(1), 129–150.

Shubik, M. (1982). *Game theory in the social sciences*. Cambridge: The MIT Press.

Smith, M. (1982). *Evolution and the theory of games*. Cambridge: Cambridge University Press.

Taber, C. S. (1992). POLI: An expert system model of U.S. foreign policy belief systems. *American Political Science Review*, 86(4), 888–904.

Tens, T., Ahmad, S., Moore, R., and Gage, E. (2004). Federal policy mandating safer cigarettes: A hypothetical simulation of the anticipated population health gains or losses. *Journal of Policy Analysis and Management*, 23(4), 857–872.

Vijverberg, W. P. (1997). The quantitative methods component in social sciences curricula in view of journal content. *Journal of Policy Analysis and Management*. 16(4), 621–629.

Wagle, U. (2000). The policy science of democracy: The issues of methodology and citizen participation. *Policy Sciences*, *33*, 207–223.

Walker, J. L. (1976). The curriculum in public policy studies at the University of Michigan. *Urban Analysis*, *3*, 89–103.

Warner, M. and Hebdon, R. (2001). Local government restructuring: Privatization and its alternatives. *Journal of Policy Analysis and Management*, 20(2), 315–336.

Weller, C. E. (2000). Risky business? Evaluating market risk of equity investment proposals to reform social security. *Journal of Policy Analysis and Management*, 19(2), 263–273.

Wells, J. B., Layne, B. H., and Allen, D. (1991). Management development training and learning styles. *Public Productivity & Management Review*, 14(4), 415–428.

Wildavsky, A. B. (1969). Rescuing policy analysis from PPBS. *Public Administration Review*, 29, 189–202.

Wildavsky, A. B. (1976). Principles for a graduate school of public policy. *Journal of Urban Analysis, 4,* 3–28.

Winter, S. C., and May, P. J. (2001). Motivation for compliance with environmental regulations. *Journal of Policy Analysis and Management*, 20(4), 675–698.

Zagonel, A. A., Rohrbaugh, J., Richardson, G. P., and Anderson, D. F. (2004). Using simulation models to address "what if" questions about welfare reform. *Journal of Policy Analysis and Management*, 23(4), 890–901.

24 The Use (and Misuse) of Surveys in Policy Analysis

Jerry Mitchell

Once upon a time in the distant past, Neanderthals were undoubtedly crouched in a cave somewhere in present day Europe wondering if they should relocate because of a shrinking bear population. If the Neanderthal's Leviathan was inclined toward a social contract way of thinking, the early humans might have been polled about their support and opposition to the move elsewhere. The results could have been used as a rationale for the risk-filled decision to budge or stay put. The reason for the eventual extinction of the Neanderthals was possibly because the populace perceived the correct policy direction, but the sovereign misinterpreted the data.

There is certainly no speculation involved in knowing that people have been formally and informally polled about different courses of action in many venues and for all sorts of reasons throughout human history. Pontius Pilate decided to put Jesus to death after taking an unsystematic survey of the local populace, President Bill Clinton decided to fib about his affair with Monica Lewinsky after his pollster told him the public would strongly disapprove of such a dalliance, and the Hungarian Parliament decided to withdraw its troops from Iraq after a poll showed 55 percent of the public favored the pull out.

It could very well be human nature for leaders and followers to question one another about what they believe or what they should do. Perhaps there is an evolutionary pressure for people to ask each other how well they have adapted or fail to adapt to environmental circumstances. After all, the pervasive propensity to gossip is nothing more than a small scale, unscientific survey that describes what other people have said and done. At the institutional level, yesterday's royal privy council and today's corporate advisory board are kindred mechanisms for eliciting opinions about particular actions. Voting is really nothing more than a state-sponsored, self-selected survey that provides a legal mandate for office holding and making laws. The fact is that people are polled about their preferences before they go to the polls and then polled again after they have been to the polls to explain why they marked their ballots one way or another.

The fascination with surveys has reached epidemic proportions. Practically every nation on the globe conducts a poll before and after the election of their leaders. In the months leading up to the 2004 U.S. presidential election, voter surveys were undertaken on a daily, if not hourly, basis by news organizations, advocacy groups, and political parties. Although politicians decry surveys and contend they are not beholden to polls, it was easy to witness the impact of surveys in the 2004 election because the two presidential candidates campaigned almost exclusively in states where surveys showed a neck and neck race. In the eccentric winner take all system of the American electoral college, there was no sense in running commercials or making personal appearances in a state where one candidate had a dominant lead according to survey research.

But it is not only in political campaigns where survey research has become popular. Viewer surveys establish which television shows survive and thrive every season, which celebrities are liked and disliked, and which commercials succeed and fail. It is a rare consumer product that has not been subjected to a marketing survey at one time or another. Customer opinion polls affect where

products are placed on store shelves and the form of advertising that appears in store windows. In fact, the University of Michigan's survey of consumer sentiment has become a leading indicator of the health of the U.S. economy. Even the determination of what is good and bad to eat is based to some degree on the longitudinal responses to questionnaires about the eating habits of some specially selected population. The extraordinary deference to surveys and the ease of their administration has led them to become a part of every school of thought, so that it is commonplace to find survey results reported in the professional journals of anthropology, psychology, sociology, education, political science, and public administration.

The use of survey research is also a part of policy analysis. Surveys are conducted to identify public needs, to discover support and opposition to policies, and to evaluate satisfaction and dissatisfaction with programs. Surveys may be employed by policy makers as the foundation for making decisions about whether to create new policies or terminate old ones, to gain a better understanding of issues, and to advocate for changing policies, programs, and services. Surveys can be applied to every stage of the policy process: to identify problems, consider the worth of solutions, determine legislative support for laws, appraise implementation difficulties, and measure outcomes. Surveys are relevant to many policy areas: the environment, social welfare, economic development, education, healthcare, civil rights, criminal justice, and foreign affairs (Christenson and Taylor 1983; Glaser and Bardo 1994; Swindell and Kelly 2000; Thompson 1997). To influence public policy, surveys are conducted by every sector—public, private, and nonprofit—from the San Francisco Zoo to the *Chicago Tribune* to the New York City Council. They can be used at every level of government: federal, state, and local. Whenever and wherever surveys are conducted, there is money to be made in putting them together and analyzing the results. In 2001, George W. Bush's administration spent nearly one million dollars alone on operations to gauge the public's reaction to alternative Social Security proposals and energy policies (Green 2002).

It does not take much noticing to notice that surveys are important part of the policy process, but surprisingly policy analysis textbooks all too often leave survey methods out of the analyst's methodological toolbox. For example, in the 499 pages of the 4th edition of David L. Weimer's and Aidan R. Vining's *Policy Analysis: Concepts and Practices* (2005) there is a mere page and half discussion of interviews, not even the inclusion of the word survey or poll in the index. Who knows the reason for this neglect, but it is surely no time to be a Neanderthal when comes to understanding how to study policies that affect the lives of countless people. The purpose of this chapter, therefore, is to examine the use of surveys in policy analysis. The first part identifies the elements of survey research, the second part provides examples of how surveys address various policy questions, and the last part examines problems with the survey research enterprise.

THE ELEMENTS OF SURVEY RESEARCH

There are four things to consider when undertaking a survey: (1) selecting the best type of survey to use, (2) developing good questions, (3) determining who should be surveyed, and (4) analyzing the results.

Type of Surveys

There are three types of surveys: telephone, in-person, and self-administered. Telephone surveys are the easiest to conduct because all that is required is a phone, phone numbers, and a caller (although large-scale telephone surveys do necessitate elaborate systems, such as telephone assisted computers and a large, well-trained staff). Interviewing people by telephone is by far the most common way of polling large numbers of people—the nation, a state, or a large metropolitan area. Telephone

surveys are advantageous because of their immediacy, standardized format, and potential for interviewers to explain questions to the respondents. However, it is impossible to reach people without telephones (the homeless, hospital patients, prisoners, etc.) and it often difficult to contact certain populations (judges, doctors, elected officials, etc.) with gatekeepers (i.e., secretary, assistant, etc.). Yet another problem is getting responses from people who employ their answering machines and caller Id systems as screening devices. Pollsters are also legally prohibited from using automated dialing equipment to call wireless numbers.

In-person surveys involve face-to-face contact between interviewers and respondents. This may involve a formatted questionnaire with a set number of responses that come one right after another or it can be unstructured with the questions evolving like a conversation between two friends. In-person surveys are not appropriate for large populations, but they are very useful when wanting to contact a select group of people in a natural setting—on the streets, in a mall, or inside a waiting room. A major advantage is to permit interviewers to explain questions to the respondents. To be done well, trained interviewers are critical because voice inflections, body language, and other physical cues can shape responses. In-person interviews are expensive and time consuming.

Self-administered surveys are distributed to respondents for completion. Surveys can be distributed in four ways: (1) sent through the mail and returned in the mail, (2) sent through e-mail or posted on a Web site and return via e-mail or by entering information on a Web site, (3) left at particular sites (on a table or counter) and either returned by mail or to the site (drop box, etc.), and (4) passed out to people as they enter or leave buildings, streets, rooms, or other venues. The advantages of self-administered surveys include anonymity for the respondents, the ability to ask sensitive questions, the potential for gaining access to difficult-to-reach populations, and the absence of interviewer bias. On the negative side, it is a difficult to obtain responses—questionnaires can be easily tossed in the trash, email can be deleted, and surveys left lying around may not be picked up. It is critical to make sure that one person does not complete more than one survey, otherwise the sample is biased. Asking good questions is extremely important because the interpretation of questions is left to the respondents.

Questionnaires

Surveys are all about questions. The conundrum is that questionnaire construction is more of an art than a science. There is no exact prescription for how any question should be asked in surveys, although there are books that provide some guidelines for asking questions, such as Peter M. Nardi's *Doing Survey Research* (2003) and Don Dillman's *Mail and Internet Surveys: The Tailored Design Method* (2000). Sometimes questions from previous surveys are repeated, but in most instances questions are crafted ad hoc from one survey to the next. Two general kinds of questions can be posed: (1) close-ended questions that provide a set of response categories for the respondents to complete, and (2) open-ended questions that allow respondents to write in their responses.

Survey questions operationalize variables. An independent variable is one that explains a behavior, attitude, or need. For example, partisan affiliation may be used as an independent variable to explain support or opposition to some policy. A dependent variable is what is being explained or accounted for. Some typical dependent variables include policy satisfaction, the use of services, and the support of public programs.

Both independent and dependent variables have different values or properties with them. For instance, age can take different values for different people or for the same person at different times. Similarly, country of origin is a variable because a person's country can be assigned a value. There are two traits of variables that should always be achieved. Each variable should be exhaustive, it should include all possible answerable responses. For instance, if the variable is "religion" and the

only options are "Protestant," "Jewish," and "Muslim," there are quite a few religions that haven't been included. The list does not exhaust all possibilities. Since it is not possible to list all possibilities with some variables, it typical to explicitly list the most common properties and then use a general category like "Other" to account for all remaining ones. In addition to being exhaustive, the properties of a variable should be mutually exclusive, no respondent should be able to have two attributes simultaneously. While this might seem obvious, it is often rather tricky in practice. For instance, it would be inappropriate to represent the variable "employment status" with the two properties "employed" and "unemployed." The problem is these attributes are not necessarily mutually exclusive—a person who is looking for a second job while employed would be able to check both attributes. The solution may to have another category "employed but looking for a job" or to have the respondent check all that apply.

Survey questions can be nominal, ordinal, or interval level measures. A nominal level measure is one that contains distant categories without any ordering. For example, if a survey asked if a person owed or rented their home. An ordinal measure is one that has a set of ordered categories. Age could be measured by a series of ordered ranges, such as from eighteen to thirty, thirty-one to forty, and so on. An interval level measure is one where every value is its own category. An example is asking an open-ended question that requires the respondent to write in the number of years they have been employed. Each response would be its own value. The level of measurement of the questions is important because it determine the kind of statistical analysis that can be performed.

There are many additional items to consider when constructing a survey instrument (Miller and Miller-Kobayahsi 2000). Respondent must be told how to answer questions and there should be a statement about whether the survey is confidential or not. Most surveys start off with questions that are relatively easy to answer, followed by more difficult questions. Demographic questions (income, age, residence, etc.) are usually posed at the end of a survey. Typically, survey researchers want to obtain an intensity of feeling in their questions, so that they would not ask if someone were satisfied or dissatisfied, but rather they would inquire whether an individual was very satisfied, somewhat satisfied, somewhat dissatisfied, or very dissatisfied. Questions should not be biased or leading. They should be easy for the respondents to understand, which requires the analyst to carefully to match questions to the units of analysis. This is one reason that surveys should be pilot tested before they are actually administered.

Respondents

There are two approaches to deciding who to survey: (1) the entire population of interest, or (2) a sample of the population. When there is a small population, everyone is usually surveyed. For instance, if one were surveying twenty-five juvenile offenders about their opinions of an alternative-to-incarnation boot camp they had just completed, then all twenty-five participants would be surveyed. There would be no need to sample them. When there is a larger population involved, then it is worthwhile to engage in sampling, that is, to draw a subset of the population. There are two types of samples: probability and non-probability.

A probability sample is one in which names are drawn from a population whose size and characteristics (such as gender, age, residence, etc.) are known. In other words, there is means to know statistically whether the sample is representative of the population. In a probability sample, it is possible to calculate a sampling error—the difference between the sample statistics and the true parameters of the population. Sampling error is a function of the number of respondents—the larger the number of people from whom data are collected, the smaller the sampling error (and, of course, the higher the cost of the survey). A survey of one thousand respondents would have a sampling error of ± 3.1 percent, while in one with two hundred respondents the sampling error

would be ±6.9 percent. Random assignment is the most common form of probability sampling, which involves giving every subject in a population the exact same chance of being selected. Another type of probability sample is a systematic sample, which involves selecting names or items from a population list at set intervals (e.g., every tenth person).

A non-probability sample is one where names are selected from a population whose size and characteristics are unknown. For example, if the Chicago Transit Authority wanted to survey its riders it would know there is a population of riders, but it would not have a master list from which to draw names. In non-probability sampling the effort is to estimate whether the sample is representative of a population that is known to exist, but whose exact parameters are unknown. To construct a representative distribution of respondents, there may be a purposive effort to obtain responses according to particular categories, such as gender, ethnicity, or occupation.

Whether it is a probability or non-probability sample, a survey researcher endeavors to have a large enough sample size to approximate the population, to have a response rate above 50 percent, and to make sure that all of the questions in the survey instrument are answered. The quality of a sample is dependent on the sample and how it will be used. If a state were considering the value of creating a new enterprise zone program and wanted to know how well it has worked in other places, it might be good enough to have a sample of the experiences of nearby states in using enterprise zones. Someone from a think tank examining the perceptions of enterprise zones in the nation would probably want a sample of American states from every region of the country.

Data Analysis

Surveys yield numbers. The irony is that subjective questions produce objective statistics. Every question in a survey is a univariate analysis that may be presented, depending on the format of the question, as a frequency distribution or measure of central tendency. Bivariate statistics depict the relationship between two questions. Multivariate statistics are about the relationship among two or more questions, which often involves the use of regression analysis. In other words, a survey assessing support for school vouchers could indicate how many of the respondents supported or opposed vouchers (a univariate analysis). It could also show whether Republicans or Democrats were more or less likely to support school vouchers (a bivariate analysis). And it could point out whether support or opposition to vouchers was affect by one variable more than others, such as partisan affiliation, gender, residence, or income (a multivariate analysis).

There are many techniques for determining the accuracy of survey results, which can be calculated using a statistical software package. For example, the Chi Square statistic measures the significance of bivariate relationships between nominal level variables while correlation coefficients measure the strength of the relationship among multiple interval level variables. Another statistic is Pearson's r, which is a measure of the strength of association between two interval level variables. The type of statistic used to assess the value of relationships is dependent on how questions are measured, the sample size, and the audience for the analysis. Complicated statistical discussions may be more appropriate for scholarly readers than for policy makers or the public.

Survey data can be presented in a narrative or in graphs and tables. If tables are used, it is important that enough information is presented for easy interpretation, but not so much information that comprehension becomes difficult. Tables should have a descriptive title, all variables and their corresponding categories should be clearly labeled, the independent variables should be listed in a column and the dependent variable should be listed along the row, statistical measures should be listed at the bottom of the table, and the number of cases used in the analysis should be indicated. After a conclusion or recommendation section, it is common for a policy report to contain an appendix which includes the survey instrument with the responses to each question.

THE USE OF SURVEYS

There are several ways surveys are used to examine public policies. Three uses are illustrative: (1) assessing the need for policies, (2) understanding the support and opposition to solutions, and (3) evaluating the responsiveness of policies to individuals and groups.

Need Assessment

Policy analysts commonly assess the need for policies among various segments of the public. How are policy makers supposed to know that policies should be adopted if they don't know what is needed? Although need is a somewhat ambiguous concept that can vary from one person or group to another, a straightforward way to understand need is to ask people what they need, letting them self-define the concept. Once the level of need has been assessed for a particular population, then a more intelligent discussion of program planning can be instigated. Ideally, policy makers and policy advocates seek to develop a service or intervention to the help the population achieve or approach a satisfactory state (Posavac and Carey 2003).

An example of a need assessment is a survey conducted by the New York City Department of Small Business Service to determine the need for a business improvement district (BID) in a local neighborhood. A BID is a professionally-managed enterprise whose purpose is to improve a locale using funds from mandatory special taxes or fees paid by property and/or business owners in a legally designated area. The issue is whether or not a BID is needed. To determine need, a survey is distributed to all of the businesses and property owners asking them to indicate whether they agree or disagree about several neighborhood conditions, such as dirty streets, pick pocketing, deteriorated facades, and retail vacancies. When the surveys results show overwhelming agreement on the severity of the problems in an area, the city council has more of a reason to approve the establishment of a BID. In fact, nearly all of the New York City's forty-seven BIDs have been established after surveys found businesses believed they were needed.

Opinion Polling

It is common to assess the level of support and opposition to alternative solutions in the policy process. Anyone and everyone can be involved in the assessment of solutions, including elected officials, public administrators, policy advocates, and journalists. Studies are done all the time to discover opinions about limiting abortion, privatizing Social Security, installing charter schools, or constructing mass transportation systems. In effect, surveys become a kind of plebiscite on the worth of policy options. If most people support some alternative, then that gives it credence, no matter whether it will or will not be effective. Conversely, if there is general opposition to an alternative, then that make may make an alternative less worthwhile, even though it might have a great chance of succeeding.

There is no more important example of how surveys are used to measure the support and opposition to public policy than in the decision of cities to build sports stadiums. Every city where new baseball, football, basketball, or multipurpose stadiums have been considered, there have been polls undertaken to diagnose the views of city residents and elected officials. These surveys are conducting by citizen groups opposed to publicly financed stadiums, business groups in support, and local media interesting in a more balanced assessment. Generally, positive survey results can give a stadium proposal the aura of respectability and negative results can make it extremely difficult to go forward. In 2001, a proposal to construct a publicly financed stadium in Minneapolis-St. Paul was seriously affected by a public opinion survey conducted by the *St. Paul Pioneer Press*. In

a front page story, the paper reported that public financing for sports facilities was an unattractive proposition among residents in the Twin Cities. Based on a telephone survey of 406 residents, 62 percent of likely voters polled in St. Paul and 71 percent of those queried in Minneapolis opposed any significant public financing for a new ballpark for the Twins. Subsequent to the survey, the stadium proposal was rejected by the city council. Although the survey was not the only reason for its demise, it was certainly a major factor

Impact Assessment

Surveys are also conducted to assess policy outcomes. People may be surveyed about whether they are aware of a public advertisement, if they have every used a revamped service, or if they are satisfied or dissatisfied with a new or ongoing program. The premise is that the capacity of a political system to respond to the preferences of its citizens is central to democratic theory and practice. From a democratic perspective, it may not really matter if a policy is effective or efficient, but instead the issue is whether or not it satisfies some segment of the public according to the results of a survey.

The evaluation of the Drug Abuse Resistance Education (D.A.R.E) program is a good example of how surveys trump other methodologies in the assessment of impact. The D.A.R.E. program involves specially trained, uniformed police officers giving lessons to elementary school students (typically, eight to twelve year olds) on how to resist drugs. By employing law enforcement officers to teach the curriculum, D.A.R.E. brings the firsthand accounts of the officers' experiences from the street to the classroom. The lessons provide factual information about drugs, with an emphasis on gateway drugs (marijuana, alcohol, and tobacco), and teach refusal skills through role-playing and other techniques. When it comes to evaluation of D.A.R.E, cost-benefit studies have consistently found it to be inefficient, and quasiexperimental designs have concluded it is not that effective in preventing young people from using drugs (Lynman et al. 1999). Nonetheless, the program survives in school districts because surveys consistently find parents, teachers, administrators, and students are satisfied with its performance. For example, a 1995 survey of 1,800 parents, teachers, and D.A.R.E. graduates in Illinois found the program was valuable and worth maintaining. Over 92 percent rated it "very good" or "good." This impact assessment is reported on the D.A.R.E Web site (2004), which, when combined with other similar surveys, is a rationale for the program's continuing presence in public schools.

THE MISUSE OF SURVEYS

The fact that surveys are used all the time does not mean they are perfect. Surveys have three problems: (1) surveys are frequently completed that are methodologically flawed, (2) surveys are regularly conducted that are politically biased, and (3) surveys are used inappropriately as a substitute for other forms of democratic engagement.

Survey Flaws

It is not easy to create the perfect survey, perhaps it is impossible. It is all together too easy to find surveys with unrepresentative samples comprised of a small number of people who choose to participate, abysmally low response rates, highly ambiguous questions, ill-defined words in the questionnaire, responses to complex subjects limited to yes and no answers, and statistics that provide percentages, but not the actual number of people who responded to the questions. The fact is that anyone can conduct a survey, without any expertise whatsoever, and there is no survey police to

hall bad researchers away. The penalty for poor research is to critique the analysis, which happens only occasionally, or to ignore the results, which happens all the time.

An example of a poorly constructed survey is a needs assessment conducted by the Los Angles Downtown Women's Action Coalition in 2001, the purpose of which was to understand the problems confronting homeless women on Skid Row. One question asked, "Overall, how would you rate the treatment you received from the staff of the various missions, shelters, and social services agencies of the Skid Row area?" The response categories were: (1) very good, (2) good, (3) average, (4) poor, (5) very poor, and (6) no opinion. The problem with the question is that it is actually three questions: one about missions, another about shelters, and third about social service agencies. The question also does not define what is meant by staff and it assumes the respondents are in agreement on where Skid Row is located. In addition, the response categories are indistinctive; is it really possible to distinguish between very good and good, or very poor and poor? The survey sample was 399, but no effort was made to show whether it was representative of the population of homeless women. The survey was completed on only one day in the summer, so it is impossible to know if there are were any seasonal variations in the opinions of the women. Simply put, the survey had many methodological flaws, although that did not stop the Coalition from publicizing the results and citing them in policy-making forums. Perhaps the methodology was not that important because the results confirmed the Coalition's advocacy work on behalf of homeless women in Los Angles.

Survey Bias

Surveys can be methodologically sound, yet biased. A tendency in policy analysis is for surveys to be created for no other reason than to rationalize, advocate, and attack public policies. In other words, surveys in policy analysis are often more political instruments than scientific endeavors. Does anyone really believe that someone with a conservative ideology would ever produce a survey that shows parents are not satisfied with school voucher programs? Has the American Association of Retired Persons (AARP) once presented statistics showing a general opposition for seniors to buy cheap drugs in Canada? Would the Sierra Club really undertake a survey to prove that most people do not believe sprawl is a major problem? It is obvious why President George W. Bush in 2004 cited a survey that showed 70 percent of Iraqis supported his policies and equally obvious why he rejected a survey that found 70 percent of Europeans were opposed to his policies.

The space between objectivity and subjectivity is ephemeral. For example, the New Haven, Connecticut Town Green District conducted a survey in 1999 with 900 surveys mailed or hand delivered to property owners and businesses, resulting in 131 responses for a response rate of 15 percent, which the district proudly noted in a newsletter was a 173 percent increase in responses over the previous year's survey. To the question—"Are you generally pleased with the impact that the Town Green has had on downtown?"—71.8 percent said "yes" and 6.9 percent said "no." To the question, "Do you see/feel a positive change downtown since the creation of the district in January 1997?"—70.2 percent said "yes" and 7.6 percent said "no." It stretches common sense to think the Town Green District would have ever conducted a survey that found over 70 percent in the no category for either of these questions. It seems apparent that this survey was more about advocacy than an empirical description of public opinion.

Undemocratic Surveys

A final issue in survey research is the presumption that the best measure of any public policy is to have a set of responses to survey questions. In effect, responding to surveys has become a substitute for other forms of forms of democratic engagement—attending public hearings, writing letters to

public officials, and voting in elections. A survey is one of a few ways that citizens can express their views about alternative garbage disposals methods in a locality or a proposal in a state to reduce a budget deficit by issuing bonds and reducing services. The assumption is that surveys are a corrective to the influence of rich elites and professional interest groups in the policy process. Other than voting, surveys are one of the few opportunities for the disadvantaged and people with busy lives to analyze and shape public policy. There is even a sense that the act of being interviewed might reduce a citizen's feeling of alienation from politics and government (Benson 1981; Web and Hatry 1973).

One problem with the use of surveys to reflect democracy is the low cost and benefits to those being interviewed (Berinsky 2004). Respondents do not contact pollsters, but instead pollsters and their political sponsors assume the costs of participation by contacting people and mobilizing them into a limited form of political action. But this only half of the equation, answering polls is also a low-benefit activity. Respondents are no better off at the beginning of an interview than at the end. In effect, to use surveys in the policy process is to rely on the lowest common denominator of democratic participation.

Another problem is that not all respondents react the same to questionnaires. Consequently, surveys tend to reinforce social inequality. Few surveys are multilingual and most require the respondents to be familiar with processing bureaucratic information. All kinds of people are excluding from survey research, such as children, people in institutions (prisons, hospitals, etc.), and individuals who for one reason or another lack the time or interest to answer survey questions. When it comes to telephone interviews, households without phone service are excluded, thereby devaluing the opinions of those with lower incomes, less stable jobs, and fewer group and community attachments (O'Sullivan, Rassal, and Bermer 2004). It is doubtful that a telephone survey on housing policy would be meaningful if it excluded those without telephones.

Polls tend to include mostly people who like putting forth their opinions about issues. There is an argument that understanding the public interest should be more than counting the opinions of individuals who enjoy giving their views on everything and anything. In fact, one of the ideas of representative democracy is that elected officials and public administrators have a responsibility to understand the silent majority. The intent of surveys is surely not to force people to have opinions. There are many policy issues when people do not have enough information to make an informed assessment. Most Americans have heard of the No Child Left Behind Act, for instance, but nearly seven in ten say they don't know enough to form an opinion about the educational initiative of the federal government.

A final problem is that surveys tend to boil complicated issues down to the level of platitudes, catch phrases, and easy to answer questions. Is it really good for democracy to think about public policy in the most simple of terms? Simplistic surveys may do nothing more than produce a convoluted mishmash of ideas and opinions that don't indicate anything other than most people are confused. Consider, for example, a *Harris Poll* conducted online in 2002 among a nationwide sample of 2,118 adults. The data was weighted to be representative of all U.S. adults. The poll found:

- Almost everyone, 93 percent of all adults, support "the United States continuing to fight . . . the war on terrorism in order to kill or capture those who planned or supported the attacks . . . on September 11th."
- When it comes to U.S. support for other countries fighting against their terrorists, the public is much more equivocal. Modest majorities favor U.S. support for Israel (63%) and Britain in Northern Ireland (56%). The public is almost equally divided as to whether the U.S. government should support the Indian government, the Russian government or the Spanish government against those attacking them in Kashmir, Chechnya and the Basque region. And most people oppose U.S. support for the Chinese government in Tibet (68%) or for "undemocratic, totalitarian or military dictatorships" (64%).

- A 58 percent majority of all adults believes that "the use of bombs and guns against . . . governments that do not give their people the right to decide their own future by free democratic elections" can be justified.
- A 57 percent majority of the public says they "think of people fighting to overthrow dictatorial, military or undemocratic governments" as freedom fighters, and only 11 percent think of them as terrorists.
- When the government is bad enough, almost everyone thinks that the use of bombs and guns against the government is justified. When told of the attempt by German officers to bomb and kill Hitler during World War II, fully 89 percent of adults say that "it is morally justified to kill people if you have no other way to fight against a really bad government or leader."

The conclusion of this poll is that the public is confused about what is and is not terrorism. It is unclear how this information helps policy makers or, for that matter, how it helped the people who responded. Terrorism is a complicated subject that requires thinking about many events in human history, weighing the advantages and disadvantages of various forms of participation in different contexts, and understanding a wide range of human behaviors. It is difficult to comprehend how simplifying reality through a series of questions that yield conflicting results will ever contribute to the improvement of public policy, one of the obvious goals of policy analysis.

CONCLUSION

The fact that there are problems with the use of surveys does not mean they should be removed from the policy process. What is needed is more careful consideration of how they are used and a better sense of the ways they can be misused. Simply put, knowledge is key to the key to good utilization. In this regard, educators must spend more effort to examine surveys in policy analysis textbooks and in college classrooms. Academics should be especially cognizant of how they report survey results in professional journals because they are implicitly setting standards for how to judge the worth of surveys conducted in the world of politics and administration. When presenting the findings of survey research, it is incumbent on everyone to be thorough in the description of their methodology. Thoroughness may go along way to eliminating mistakes, exposing bias, and indicating how the results should be used. A survey report that details how questions were asked, notes all aspects of the sampling procedure, and explains the statistics analysis will be much more likely to be utilized and understood. And when the elements of survey research are followed closely and the problems with survey research are avoided, the use of surveys in policy analysis will be less Neanderthal and more likely to lead to good public policy.

REFERENCES

Benson, Paul R. (1981). Political Alienation and Public Satisfaction With Police Services. *Pacific Sociological Review* 24(1): 45–64.

Berinsky, Adam J. (2004). *Silent Voices: Public Opinion and Political Participation*. Princeton, NJ: Princeton University Press.

Christenson, James A., and Gregory S. Taylor. (1983). The Socially Constructed And Situational Context For Assessment of Public Services. *Social Science Quarterly* 64(2): 264–274.

Dillman, Don. (2000). *Mail and Internet Surveys: The Tailored Design Method.* New York: Wiley.

Downtown Women's Action Coalition. (2001). *Downtown Women's Needs Assessment: Findings and Recommendations.*

Glaser, Mark A., and John W. Bardo. (1994). A Five Stage Approach For Improved Use of Citizen Surveys in Public Investment Decisions. *State and Local Government Review* 26(3): 161–172.

Green, Josuha. (2002). The Other War Room: President Bush Doesn't Believe in Polling—Just Ask his Pollsters. *The Washington Monthly*, (April).

Harris Poll. (2004). http://www.harrrispoll.com.

Kelly, Janet, and David Swindell. (2002). A Multiple-Indicator Approach To Municipal Service Evaluation: Correlating Performance Measurement and Citizen Satisfaction Across Jurisdictions. *Public Administration Review* 62(5): 610–621.

Lynam, Donald R., Richard Milich, Rick Zimmerman, Scott P. Novak, T. K. Logan, Catherine Martin, Carl Leukefeld, and Richard Clayton. (1999). Project DARE: No Effects at 10-Year Follow-Up. *Journal of Consulting and Clinical Psychology* 67, No. 4 (August): 590–593.

Miller, Thomas L., and Michelle A. Miller-Kobayahsi. (2000). *Citizen Surveys: How To Do Them, How To Use Them, What They Mea*n. Washington, DC: International City Management Association.

Nardi, Peter M. (2003). *Doing Survey Research: A Guide to Quantitative Methods*. New York: Longman.

O'Sullivan, Elizabethann, Gary R. Rassal, and Maureen Berner. (2004). *Research Methods for Public Administrators.* New York: Longman.

Posavac, Emil J., and Raymond G. Carey. (2003). *Program Evaluation: Methods and Cases*. Upper Saddle River, NJ: Prentice Hall.

Project D.A.R.E. (2004). http://www.dare.com.

St. Paul Pioneer Press. (2001). Little Support for Public Financing for Twins' Stadium, (November 1), Internet edition.

Swindell, David, and Janet Kelly. (2000). Linking Citizen Satisfaction Data to Performance Measures. *Public Performance & Management Review* 24(1): 30–52.

Thompson, Lyke. (1997). Citizen Attitudes about Service Delivery Modes. *Journal of Urban Affairs* 19(3): 291–302.

Weimer, David L. Weimer, and Aidan R. Vining. (2005). *Policy Analysis: Concepts and Practices.* New York: Prentice Hall.

Webb, Kenneth, and Harry P. Hatry, H. (1973). *Obtaining Citizen Feedbac*k. Washington, DC: The Urban Institute Press.

25 Social Experiments and Public Policy

Caroline Danielson

INTRODUCTION

Social experiments randomly assign people (or sometimes sets of people, i.e., neighborhoods or communities) either to a group that is subject to one or more policy "treatments" or to one that continues to be subject to the prevailing policy norm ("controls"). For example, a social experiment might test the efficacy of a welfare-to-work program by randomly assigning welfare applicants to the new program (perhaps an intensive, coached job search combined with the provision of services like transportation assistance and subsidized child care) and to the old standard, which leaves the initiative to find a job nearly completely up to the welfare recipient.[1]

It is standard for those who conduct experiments to make the claim that theirs is the only methodology that can with certainty isolate the impact of the program under evaluation. Social experiments alone can assure that "any differences that emerge over time in employment, earnings, or other relevant outcomes can reliably be attributed to the program" (Berlin 2002, 3). Yet this is not simply a claim that circulates in the research community. Social experiments also hold the respect of those crafting social policy (Baum 1991; Greenberg, Linksz, and Mandell 2003; Greenberg, Mandell, and Onstott 2000; Haskins 1991). Social experiments generate this respect because they appear to offer a readily-accessible, incontrovertible answer to the most pressing question in evaluation research: does X program cause Y outcome?[2]

In this chapter, I examine the key factors that make social experiments attractive to both researchers and policy makers. These features of social experiments seem worth exploring because there appears to be a consensus among researchers and policy makers that experiments constitute a gold standard in policy evaluation. To the extent that this consensus exists, it removes one obstacle to the application of social science research to policy making. Social experimentation promises to be a rigorous, straightforward arbiter among political choices—a method well-suited to the division of labor that leaves the choice of ends to policy makers and the evaluation of means to technical experts.

Such a consensus clearly does not imply that only evidence from social experiments will be used in the policy process, or even that any research at all will guide policy making. The literature describing the actual use of research in policy making is also extensive. Greenberg, Linksz, and Mandell (2003) explore the influence of social experiments in the welfare policy arena on state policy makers, and Weaver (1999) examines the role of policy research in the debates on "ending

1. I use examples from the arena of welfare policy to illustrate principles and pitfalls of social experiments; however, social experiments are also used in other social policy arenas, including crime, education, and health. For a list of major social experiments conducted in the United States, see Greenberg and Shroder (2004).
2. I focus on tests of the efficacy of a program, although experimental data can be turned to other purposes—for instance, computing cost-benefit calculations.

welfare as we know it" that occurred in the 1990s. Aaron and Todd (1979) reports on the influence of earlier social experiments on policy. For examples of other research examining more generally the use of social science research in policymaking (see Danziger 2001; Haveman 1976; Hird 2005; Jones 1976; Lindblom and Cohen 1979; Rich 2001; Shulock 1999; Stone 1997; Szanton, 1981). Obviously, legislators can choose whether to request, or to use, social scientific research to evaluate policy proposals although their choices are constrained by the prevailing norms regarding the applicability of research to policy development and evaluation. This chapter will examine key aspects of the prevailing norms regarding social experiments.

A number of important aspects of social experiments have been discussed extensively elsewhere: technical issues (e.g., selective attrition, determining the effect of the treatment), the definition of treatments (are only certain types of policies tested? are programs tested in an intensive enough way?), and the practicalities of running experiments (obtaining the committed participation of agencies that implement policies, adequately getting the message across). Recent discussions include Gennetian et al. (2002), Grogger and Karoly (2005), Haveman (1987), Heckman in Manski and Garfinkel (1992), Lalonde (1995), and Orr (1999).

I focus here on unpacking two accepted aspects of experiments that make them attractive to policy makers: their ability to isolate causes and their methodological transparency. Experiments offer these virtues, but not in an unqualified way: they are not a complete recipe for policy evaluation. I also take up ethical questions that social experiments pose. I argue that the absence of a real debate over ethics is more evidence that social experiments are an established methodology from the point of view of both researchers and policy makers. My argument is in line with other discussions of the ways in which methodology can take precedence over substantive debates about the ends that democratic societies seek to achieve and the permissible means that they can use to achieve them (Fischer 1990; Fischer 2003; Stone 1993).

After sketching the history of social experiments in the next section and summarizing essentials of conducting experiments in the third, I take up the primary intellectual attraction of experiments in the fourth—their ability to isolate the program from other events that shape subjects' outcomes—and in the fifth I discuss a central political attraction of experiments—what I call their transparency. In the sixth section I review standard ethical justifications of social experimentation.

EVOLUTION OF SOCIAL EXPERIMENTATION

Greenberg, Linksz, and Mandell (2003) review the history of social experiments and *The Digest of Social Experiments* describes all social experiments conducted to-date in the United States (Greenberg and Shroder 2004). When large-scale social experiments were first proposed in the 1960s, they were a departure from the normal practice of policy research. Evidence of this is that organizational capacity had to be built to handle the new demand: the Manpower Demonstration Research Corporation (now simply MDRC) a non-profit, non-partisan research organization, with seed funding from the Ford Foundation, was purpose-built in the early 1970s to conduct the National Supported Work Demonstration (Gueron 2000; Manski and Garfinkel 1992).[3] A handful of other organizations also retooled to undertake experiments (e.g., Mathematica Policy Research, Abt Associates, Rand, organized research units at a few large universities). The scale and scope of cooperation between civilian agencies and researchers was also unprecedented.

Haveman (1987) notes that poverty research, and the organizations capable of training researchers and carrying out the research, were fundamentally shaped by the War on Poverty, which provided the funding and the federal agency loci to stimulate policy research in this field. The *Digest*

3. The Ford Foundation also funded the development of several public policy schools (Haveman 1987), presumably to develop capacity to conduct rigorous, policy-relevant research.

of Social Experiments reveals that the majority of social experiments have been conducted with poor populations as subjects in program areas that include health, employment, and education and training (see Greenberg and Shroder 2004).

The first social experiment, the New Jersey Negative Income Tax Experiment (NIT), was conducted by Mathematica under contract from the University of Wisconsin's Institute for Research on Poverty. This experiment had several treatment groups, each of which was subject to a different combination of a minimum guaranteed income and a tax rate on income earned above the guarantee. The core aim was to test whether adults would reduce their hours of work if they knew they were guaranteed a minimum income. According to an observer, it was not obvious that experimentally altering individual's incomes was ethical. Conducting the New Jersey NIT was justified on the grounds that there was no other way to obtain answers to the question of individuals' responses to a guaranteed income (Haveman 1987).

It is important to realize that experiments were first intended to be used in conjunction with simulation to provide a way of projecting the impact of a broad range of policies. According to Haveman (1987), writing after the first wave of social experimentation in the 1970s had ended, a goal of all of these experiments was to estimate structural parameters like the behavioral response (expressed, for instance, in hours of work) to manipulations of income by tax policy. That is, so long as the assumption could be maintained that individuals' behavior in response to incentives like additional income was constant across time and place and varied smoothly, an experiment that assigned groups to several gradations of tax policy treatments could be used to estimate the impact of a whole range of tax policies on hours of work.

According to economist James Heckman, however, as the NIT experiment in particular progressed, its aims grew more constrained: it came to be to compute the mean impacts of the program (Heckman in Manski and Garfinkel 1992). Instead of one or several experiments providing the raw material that would enable researchers to simulate behavioral responses to a range of hypothetical policies, an experiment would supply simply the difference in outcomes between the treatment and the control group for the policy or policies under study. This type of experiment gets called a "black-box" experiment because researchers make no strong claims about the underlying causes of the outcomes; their focus is on reporting the results of a particular policy treatment.

The scaling back of researchers' ambitions had partly to do with technical difficulties in collecting data adequate to the task of simulating responses to a range of hypothetical policies. The results of the NIT, in particular, were not as clean as expected. Apparently the implied responses to different tax policies the researchers computed relied on self-reported, and therefore incomplete, income data as well as on the experimental data, and the computations did not produce a smooth pattern of responses. But dropping this more theory-laden approach to experiments perhaps also betrays the insight that a more easily-communicated approach is more compelling to policy makers. Black-box experiments report the outcome, attribute it to the treatment, and stop there.

The goal in conducting social experiments has decidedly shifted to estimating mean impacts of the treatment (Greenberg, Linksz, and Mandell 2003). It is this way of setting up social experiments that has won the approbation of policy makers and that backs confident statements that organizations like MDRC make about the methodological rigor of experimental evaluations. As the then president of MDRC has stated, "With random assignment, you can know something with much greater certainty and, as a result, can more confidently separate fact from advocacy" (Gueron 2000, 1).

The sense that black-box experiments are the gold standard of evaluation research had developed by the late 1980s. According to those pivotal in developing the legislative language for the 1988 overhaul of the Aid to Families with Dependent Children (AFDC) program, Family Support Act (FSA), experimental evidence from a number of welfare-to-work projects that MDRC was conducting played a decisive role (Baum 1991; Haskins 1991). This was the case for a number of reasons (fortuitous timing, MDRC's ability to both disseminate results widely and in a timely fashion and to maintain a non-partisan stance), but included the absence of debate among researchers about

the outcomes and the concomitant respect that the methodology commanded. The FSA included provision for the evaluation of its effects using randomized experiments.

A few years later, a little-used provision of Title IV-A, Section 413 of the Social Security Act stating that federal requirements for AFDC could be waived by the Secretary of the US Department of Health and Human Services was exploited to allow states broadly to experiment with their AFDC programs in the early- and mid-1990s. Section 413 states that "The Secretary may assist States in developing, and shall evaluate, innovative approaches for reducing welfare dependency and increasing the well-being of minor children living at home". It continues, "In performing the[se] evaluation[s]…, the Secretary shall, to the maximum extent feasible, use random assignment as an evaluation methodology" (42 U.S.C. 613).

While obtaining approval for waivers to existing AFDC program regulations apparently became quite straightforward by the mid-1990s—the Clinton administration did not want to be perceived as obstructing states' reform efforts—the Administration for Children and Families did typically require states to perform randomized experimental evaluations of their programs, a requirement that produced a wealth of data that would not otherwise exist. Forty-three states obtained waivers between January 1993 and August 1996, although not every state actually implemented its waiver program (Boehnen and Corbett 1996; Gordon, Jacobson, and Fraker 1996). With this impetus, a large number of social experiments was initiated in the 1990s.[4] One might say, then, that the early- and mid-1990s marked a high point in policy evaluation because of the widespread use of experimental methodology.

Although the pace of experimentation has since slowed in the welfare policy arena, commentators continue to call for experimental evaluations of policy proposals newly on the national agenda. For example, a policy brief published by the Brookings Institution endorses experimental evaluations of government programs to encourage marriage as a means of defusing controversy over their appropriateness—if programs that encourage couples to marry raise marriage rates, then, presumably, concerns about intervening in individuals' private lives will diminish (Haskins and Offner 2003). At the same time, the case for experimental evaluation of policy proposals is building in other policy arenas. As evidenced by the language of the 2002 No Child Left Behind act policy makers are now advocating evaluations of policy proposals the field of education using experimental methodology (Glenn 2004; Mosteller and Boruch 2002).

NUTS AND BOLTS OF EXPERIMENTS

To conduct an experiment, researchers randomly assign some members of a target group to the program under study and some to the current program. The impact of the treatment is measured as the mean difference between the treatment and control groups on relevant measures (e.g., income, educational achievement, mental health). That is, how much more (or less) income did the treatment group earn at the end of the study period than the control group did? Or, how much higher (or lower) did the treatment group score on a standardized test administered to both groups?

Internal validity is the core methodological strength of experiments. Assigning members of a target group at random to treatment and control comparison groups ensures that they are statistically equivalent on both measured and unmeasured characteristics. Since adjustments can be made for differences on measured characteristics, the problem that other research methods face is their inability to methodologically rule out systematic differences between nonexperimental comparison

4. The evaluated programs include the Minnesota Family Investment Program (MFIP), Florida's Family Investment Program (FIP), Vermont's Welfare Restructuring Project (WRP), Arizona's EMPOWER program, Connecticut's Jobs First program, Iowa's Family Investment Program (FIP), and the Indiana Manpower Placement and Comprehensive Training Program (IMPACT).

groups on *unmeasured* characteristics.[5] Experiments make it possible to confidently assert that there are no differences (statistically speaking) between experimental comparison groups on unmeasured characteristics. Any differences between groups measured subsequently can therefore be confidently attributed to the treatment, within the bounds of certainty provided by statistics.

To insure the internal validity of experiments, researchers must successfully randomize participants between the test and the standard programs. This involves developing a protocol for initial randomization that is straightforward and not susceptible to manipulation by those implementing the protocol. It further involves ensuring that members of the control group do not cross over and obtain the program reserved for the treatment group. It also requires that members of the treatment group realized that they were subject to different program rules than the control group. For more extensive treatments of the practicalities of conducting experiments, see Boruch (1997), Hausman and Wise (1985), and Orr (1999).

It is important to be clear about what the outcomes that can be measured experimentally are. The experimental outcomes that can be measured depend on the point at which randomization occurred. For example, if some welfare recipients are assigned to have a limit on the number of months they are eligible to receive cash assistance and some are not, then the experimental outcomes that can be measured are in relationship to *exposure* to a time limit. For example, did those who were subject to a time limit find a job sooner than those who were not? Or, did they use fewer months of welfare over a particular period of time? The effect of the time limit on the income and well-being of those who reach it in the experimental group is not an experimental outcome, since the two groups were not randomly assigned to *reach* the time limit. It is possible to compare subgroups defined by initial characteristics of the treatment and control groups because the groups are (stastistically) identical on those measures. For example, it would be possible to compare the differences between long-term welfare recipients assigned to the groups subject and not subject to the time limit.

A related point is that experimental outcomes are measured as differences between those assigned to the program and those not assigned to it. The effect of the new program is often not identical to the impact measured by the experiment because not everyone gets the program. For example, some of those assigned to a treatment group that is eligible for a range of job search services will not avail themselves of any of the services, or of only some of them.[6] Finally, the impact of the program either on participants or on those randomized to be eligible for the program is not the identical to the impact on the population if the new program were to become policy because experiments typically do not randomly assign entire target populations to treatment and control groups. The first crucial point here is that a new program may very well change the applicant pool. For example, in the presence of time limits, some of those who would earlier immediately have applied for cash assistance when they experienced a job loss might hold out for a few months, realizing that they now only have a limited number of months of eligibility for cash aid.[7] The second is that if a new program is widely implemented, it may change the broader environment in which it operates (so-called "macro" effects) in ways that a small pilot program that is experimentally tested would not. For example, a job search program, if implemented for all welfare recipients, may alter the labor market for low-income workers, thus altering the effectiveness of the job search program.

5. Researchers using nonexperimental methods can argue that, for theoretical or practical reasons, it is unlikely that the comparison groups in question differ on unmeasured characteristics.
6. Random assignment experiments, under certain assumptions, have a built-in instrumental variables estimator that can be used to estimate the average effect of the treatment on the treated (Angrist et al., 1996; Gennetian et al., 2002).
7. While it would be possible to randomly assign a sample of the entire target population to treatment and control groups, it would be more expensive (and in many cases prohibitively so) because a large enough sample would have to be randomized in order to detect the effect of the treatment. The size of this sample would depend on the expected rate of application to the program among members of the target population.

CONCEPTION OF CAUSALITY

Here is a stripped-down version of the core question to which policy makers seek an answer when they commission a policy evaluation: If we implement X program, will Y outcome result (or, in the case of a program already implemented: Did X program produce Y outcome that we envisioned)? Policy evaluation is fundamentally a testing of means. Simplifying the real complexities of the process of policy making, one can say that policy makers seek to achieve an end. The ideal evaluation of a policy would answer the question, does one particular means as compared to another advance us toward that end?

Here is the question that social experiments address: On average, there was (or was not) a statistically significant difference (at conventional levels) between the outcomes of treatment group T and control group C on measure M (of outcome Y) in an experiment in which X program was tested. For example, if policy makers want to know whether a welfare-to-work program that emphasizes quick immersion into a process of searching for a job (X) improves child well-being (Y), researchers would design an experiment that randomly assigned some (T) to participate in a sequence of job search activities and others not (C). Child well-being (Y) might be measured, among other things, by surveying parents about problem behaviors their children might be exhibiting (M).

Is the question that policy makers implicitly pose identical to the question that researchers address? The central difference between the two questions posed above is generalizability. It seems clear from the way that the first query above is framed that policy makers are interested in a general result, or something resembling law-like behavior. If program X is funded, then Y outcome will always (or usually) obtain. But experiments tell us nothing directly about law-like behavior. Their methodological soundness comes exactly from their internal validity. That is, experiments are a powerful means of attributing the impact of the intervention, and not other factors, to the outcomes observed by researchers. Experiments accomplish this by posing a counterfactual: what would have happened had the program *not* existed? Thus researchers use experiments to identify causes using the evidence from unique occurrences, rather than that obtained from observing regularities or from logical deduction.[8]

Given policy makers' interest in the more general question, the natural inclination is to generalize. Thus a natural slippage occurs: researchers and policy makers treat the experiment as predictive of outcomes in other times and places that are "similar enough." But what counts as similar enough? What would the outlines of an argument that generalized from one particular experiment look like? There are two key elements: (1) identify the most important behavioral mechanisms that produced the result, and (2) identify key features of situations that make them enough like the experimental situation so that individuals placed in those like situations will interact with the context in the same manner that the experimental subjects did.

Because experiments take a black-box approach, they do not address the behavioral responses that the program may have induced (although researchers can and do use other methodologies to understand such mechanisms). And unlike researchers conducting laboratory experiments, those carrying out social experiments do not control the context in which the treatment and control group programs unfold. In this sense, they cannot rigorously specify the context. There is at least one strong reason to believe that experimental situations are exceptional: those who are "treated" are not blind to their situation, and those who administer the treatment often know the circumstances of the experiment—this is a crucial difference between double-blind medical trials where both treatment and control groups are treated and neither researcher nor subject knows who received the treatment and social experiments.

Those interpreting experimental impacts must make additional inferences in order to generalize beyond the particular instance, and they must do so on grounds other than the soundness of the

8. Max Weber (1949) developed this conception of causality.

internal validity of the experiment.[9] The causal question to which policy makers seek an answer thus differs crucially from the question that researchers answer by conducting a social experiment.

METHODOLOGICAL TRANSPARENCY

Perhaps experiments must be interpreted with caution because they do not unpack causal mechanisms and because conclusions drawn from them do not extend in a straightforward fashion to programs put in place in other contexts. But as freestanding exercises, experiments have the virtue of employing a methodology that is more readily grasped than other evaluation methodologies. In addition, experiments are attractive because they promise to sidestep the debates of "dueling witch doctors" that heighten the politicization of policy debates: when technical experts disagree, it undermines the credibility of the policy proposal (Baum 1991).

The promise that social experiments make of a more immediate, incontrovertible truth than other research methodologies offer appears to rest on two factors. First, grasping the essentials of social experiments seems to require no arcane technical training inaccessible to policy makers and their advisors. Second, and relatedly, the outcomes of experiments are not murky: experiments reliably allow observers to sharply distinguish between programs that worked and those that had no effect on outcomes of interest.

One might complain that a key test of a social scientific methodology in the policy arena should not be its (apparent) lack of technical complexity, but this complaint would be misplaced. Garfinkel, Manski, and Michalopoulos claim that social experiments receive funding preference over basic research in the social sciences because policy makers are unable to interpret the disputes that social scientists enter into over the results of quasi-experimental research (in Manski and Garfinkel 1992). But when, for example, engineers and biologists are hired by policy makers, they produce proof of the viability (or lack thereof) of their efforts: an unmanned air vehicle that can track a highway, a fly-sized drone that collects photographic evidence. Social scientists in general face precisely the problem that they cannot produce tangible proof that social programs are working without simultaneously justifying and explaining the methodology by which they arrived at their conclusions. That is, a welfare-to-work program must be shown to be effective; it is not evident from simple observation whether the program increased subjects' hours of work or not.

There is, in fact, a large literature on subtle technicalities of experimentation. These subtleties range from the step of the program at which randomization occurs (experimental differences must be measured in relationship to this step) to the difference between intention to treat and the effect of the treatment to macro effects that experiments do not fully capture like information diffusion, norm formation and altered market equilibria.[10] These subtleties typically receive only scant attention when researchers communicate their results to policy makers (see, for example, Hamilton et al. 2001, ES-9; Beecroft et al. 2003, ii).

It is worth noting in this context that the experimental evaluations of welfare-to-work programs, among them those that so impressed the framers of the FSA in the late 1980s, have now been subject to several reanalyses that raise questions about the internal validity of the findings (Hotz, Imbens, and Klerman 2001; Walker et al. 2003). That is to say, there is debate among researchers about the outcomes of the very experiments that had such an influence on the formulation of the work-first approach in state reforms and eventually on the shape of the Temporary Assistance for Needy Families (TANF) program that replaced Aid to Families with Dependent Children (AFDC) in 1996. As is probably often the case, this scholarly debate took longer to mature than did the policy debate.

9. Other sorts of evaluation methodologies also pose problems for generalization. Manski (1995) addresses this issue in a broader sense.

10. For examples of these discussions, see Manski and Garfinkel1992. See Haveman (1987) for a discussion of problems that the first set of social experiments shared.

In fact, it is plausible that the aims of social experiments and the manner in which their results are reported is heavily influenced by the desire to communicate in a transparent way. As I described in section two, economists' original aim for experimentation was to recover structural behavioral parameters that could be used to simulate the impact of arbitrary policies. But this more ambitious aim was quickly dropped, possibly partly because it was not compelling to policy makers.

Further, it might seem puzzling that experiments are typically agnostic about the outcome. Even if theory or intuition predicts that welfare recipients who receive job search assistance should find employment at a higher rate than those who do not, researchers perform a two-tailed hypothesis test (i.e., that the alternative hypothesis is the difference of means is equal to zero, not the difference of means is greater than zero).[11] This unwillingness to begin from theory is perhaps more evidence that experiments are meant to be transparent, or assumption-free. Alternatively, it is possibly evidence that researchers seek to be as conservative as possible.

Thus it might be fair to say that the way social experiments have been carried out has been influenced by policy makers' need for simplicity and clarity. But it would be misleading to state that experiments are simply methodologically transparent.

ETHICAL JUSTIFICATIONS

As I noted in the second section, conducting the first NIT experiment was acceptable because it was seen as a last resort: those who proposed it and those who supported its implementation could envision no other way of testing individuals' responses to a guaranteed minimum income. Social experiments must no longer meet this severe standard. They are now presumed to be appropriate: randomized trials are ethical except in special circumstances. For an example of the standard defense, see the statement by the then-president of MDRC (Gueron 2000): In short, resources can generally be presumed to limited, and as long as more people are potentially eligible for the program than can actually be served, random assignment is a fair way of allocating scarce resources. More fundamentally, random assignment is a means of determining whether programs benefit target populations or not. Both individual and social ends can be better achieved if only successful programs are pursued with government dollars.

Although it is apparent that not every situation of policy interest is susceptible to experimentation for ethical reasons, versions of random assignment seem to be. For example, it is not possible to imagine assigning children to parents, or education to children, even though it is vitally important to know how much difference family background, and how much education, makes to children's achievement. But it is possible to contemplate assigning parents to programs that increase their likelihood of developing positive relationships with one another and their children (Haskins and Offner 2003; Dion et al. 2003). While proposed programs are not identical to "assigning children to parents," the shaping of choices that these programs, if successful, would have in effect imply that some children will have relationships with parents that they would not otherwise have had.

It appears that they are presumed to be ethical because of aspects of the methodology employed. Researchers point out if resources are limited so that only a subset of applicants can be accommodated in the program, then random assignment—the core distinctiveness of social experiments—is a fair way of distributing the opportunity to participate, and is more fair than the most likely other means of allocating it (e.g., first come, first served). This argument can be challenged. Researchers must ensure that the treatment and control groups are large enough to produce reliable estimates of the impact of the program. Depending on the size of the program being evaluated, they may warn sponsors that evaluation sites will need to ramp up recruitment efforts in order to enroll enough subjects to randomize (see, e.g., Dion et al. 2003). In such situations, everyone who sought the service or program being evaluated could be accommodated, and it is the experiment that produces

11. There are also standard phrases to repeat here—Type I error, Type II error, replication—I will just note them here in passing. They also bear on the reliability of the distinction made in crucial ways.

the need to deny some access.

In cases where programs are mandatory for all applicants, or are open to all who request it, random assignment can be thought of as a fair way of assigning recipients to programs of unknown efficacy. That is, if it is unknown whether the old or the new program produces better outcomes, a social experiment can determine whether the new program should continue. Once the experiment has run its course, the knowledge that it produces will benefit all future applicants. Note that if experiments are the gold standard, then it is tautologically true that the program's effects are not known—at least not with any credibility—in the absence of one or several social experiments conducted to establish the efficacy of the program.[12]

It is also the case that, in the United States, the provision of social supports are typically not seen as rights and poor people are not taken to be a group that requires special protections. Just as the government can grant or withhold tax relief at will, denying an individual access to a program to which she does not have a strong claim, even if comparable others do have access to the program, poses a weak ethical dilemma. This ethical dilemma is further weakened by an individualist ethic. Social experiments do not deny individuals access to services that they might desire because those in the control group can often, through their own initiative, acquire the education or job search assistance that the experimental program provides to some.

In these ways it has become easy to justify random experiments as an evaluation tool for any program that Congress or a state legislature might decide to authorize. Randomization to be eligible for (temporary) income supplements, to have time-limited welfare benefits, and to receive job search assistance have all recently passed muster. So while on the ground there are apparently ethical qualms about assigning participants to treatment and control groups that are serious enough to cause agencies to refuse to participate in experiments—see , for example, Gueron (2000)—the research community's justifications for experimentation makes clear that researchers see no serious ethical barriers to randomization in a broad range of instances. To an extent, social experiments are even treated as establishing a baseline for the ethical evolution of social policy: random assignment is a fair means of allocating scarce resources and experiments can tell us which programs help, and which harm, target populations.

There is a related question that is a natural follow-on for those who promote social policy evaluation via experimentation: Why does a society aim to understand the effects of programs on outcomes? A straightforward answer is that its members seek to improve the lot of disadvantaged groups. Since social programs can appear promising without in fact producing intended effects, evaluating their effects again makes sense, and social experiments are the most reliable means of evaluating their effects.[13]

In this sense, the ethical question has been turned on its head: Is it ethical to assign (or even invite) individuals to participate in programs *without* knowing whether the treatment has the intended effect? After all, social choices are about achieving ends. While they may limit permissible means, programs like TANF are primarily aimed at achieving certain outcomes (broadly speaking, ameliorating the lot of children in poor families). Social experiments are a means of achieving social ends by helping to determine which programs further social goals Without delving into mechanisms for social choice and their justifications, it appears to be at least possible to assert that research methodologies that seek to isolate the impact of programs on participants advance social choice-making by advancing the achievement of social ends like that of improving the well-being

12. Because of the large number of factors outside of the control of those orchestrating the experiment, it is probably quite difficult to replicate a social experiment. All the same, the argument in favor of replication does not lose its punch: as is the case with all research that relies on statistics, experimental impacts are subject to random error. Then the questions becomes: at what point do researchers decide that they know whether the program works or not, and thus whether this ethical justification for randomization still holds?

13. See O'Connor (2001) for the claim that poverty research focuses narrowly on individual circumstances and behaviors instead of on the social and economic opportunities that these individuals face.

of low-income children.[14] Social experiments achieve particularly high marks in this regard because it is accepted that they are methodologically rigorous.

The ethical questions posed by experimentation are no longer ones that set up serious obstacles to the implementation of experiments because the technical merit of experimentation is unquestioned. A number of researchers have explored different aspects of the push toward neutral, efficient decision-making in the policy arena, often noting a link to anti-democratic practices (Fischer 1990; Fischer 2003; Stone 1993). Fischer (1990) argues that the objective of technocrats is to remove as many decisions from the political arena as possible, shifting them into the arena of administration. In the case of social experiments, it appears that placing faith in their methodological virtues has allowed policy makers to largely finesse ethical questions.

CONCLUSIONS

To recap, social experiments appear to offer three core strengths: fairness, simplicity, and rigor. Random assignment offers a fair way (perhaps the most fair way) of allocating scarce resources, no special training is required to grasp the essentials of the method, and experiments reliably isolate "the program" from other factors influencing subjects' outcomes thus informing policy makers how to further social ends. Experiments do possess these virtues, but they do not possess them in an unqualified way. I have argued that, in fact, the conclusions that can be drawn from social experiments, like other evaluation methodologies, rest on crucial assumptions, and they have important limitations. The manner in which social experimentation has evolved in the United States has reduced the apparent complexities of evaluating social policies, but it has not actually erased them. Policy makers would do well to keep these complexities and limitations in view, even as they point to the strengths of experimentation. Finally, the fact that experiments are methodologically attractive should not be a reason to sideline ethical questions.

REFERENCES

Aaron, Henry J., and John Todd. 1979. "The Use of Income Maintenance Experiment Findings in Public Policy, 1977–78." *Industrial Relations Research Association Proceedings, 1979*. Madison, WI: IRRA.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434, June): 444–455.

Baum, Erica. 1991. "When the Witch Doctors Agree: The Family Support Act and Social Science Research." *Journal of Policy Analysis and Management* 10(Fall): 603–615.

Beecroft, Erik, Wang Lee, and David Long. 2003. *The Indiana Welfare Reform Evaluation: Five-Year Impacts, Implementation, Costs and Benefits.* Cambridge, MA: Abt Associates, September.

Berlin, Gordon L. 2002. "Encouraging Work, Reducing Poverty: The Impact of Work Incentive Programs." New York: Manpower Demonstration Research Corporation, March.

Boehnen, Elisabeth, and Corbett, Thomas. 1996. "Welfare Waivers: Some Salient Trends." *Focus* 18(1): 34–41.

Boruch, Robert F. 1997. *Randomized Experiments for Planning and Evaluation: A Practical Guide.* Thousand Oaks, CA: Sage.

Brady, Henry E., Nancy Nicosia, and Eva Y. Seto. 2002. "Establishing Causality in Welfare Research: Theory Application. Unpublished Manuscript.

Burtless, Gary. 1995. "The Case for Randomized Field Trials in Economic and Policy Research. *The Journal of Economic Perspectives* 9(2, Spring): 63–84.

Danziger, Sheldon. 2001. "Welfare Reform from Nixon to Clinton: What Role for Social Science?" In *Social*

14. For a critical review of some issues in social choice literature, see Sen 1982.

Science and Policy-Making: The Search for Relevance in the Twentieth Century. D. L. Featherman and M. A. Vinoviskis, eds. Ann Arbor: University of Michigan Press, 137–164.

Dion, M. Robin, Barbara Devaney, Sheena McConnell, Melissa Ford, Heather Hill, and Pamela Winston. 2003. "Helping Unwed Parents Build Strong and Health Marriages: A Conceptual Framework for Interventions. Final Report submitted to the Administration for Children and Families, US DHHS, January.

Fischer, Frank. 1990. *Technocracy and the Politics of Expertise.* Newbury Park, CA: Sage.

Fischer, Frank. 2003. *Reframing Public Policy: Discursive Politics and Deliberative Practices.* Oxford: Oxford University Press.

Gennetian, Lisa A., Johannes M. Bos, and Pamela A. Morris. 2002. "Using Instrumental Variables Analysis to Learn More from Social Policy Experiments." MDRC Working Papers on Research Methodology. New York: Manpower Demonstration Research Corporation, October.

Glenn, David. 2004. "No Classroom Left Unstudied." *Chronicle of Higher Education* 50(38): A12.

Gordon, Anne, Jonathan Jacobson, and Thomas Fraker. 1996. "Approaches to Evaluating Welfare Reform: Lessons from Five State Demonstrations." Cambridge, MA: Mathematica Policy Research, October.

Greenberg, David and Mark Shroder. 2004. *The Digest of Social Experiments* (3rd ed.). Washington, D.C.: Urban Institute Press.

Greenberg, David, Donna Linksz, and Marvin Mandell. 2003. *Social Experimentation and Public Policymaking.* Washington, D.C.: The Urban Institute Press.

Greenberg, David, Marvin Mandell, and Matthew Onstott. 2000. "The Dissemination and Utilization of Welfare-to-Work Experiments in State Policymaking." *Journal of Policy Analysis and Management*. 19(3): 367–382.

Grogger, Jeffrey and Lynn A. Karoly. 2005. *Welfare Reform: Effects of a Decade of Change.* Cambridge, MA: Harvard University Press.

Gueron, Judith M. 2000. "The Politics of Random Assignment: Implementing Studies and Impacting Policy. New York: Manpower Demonstration Research Corporation.

Gueron, Judith M. 2003. Fostering Research Excellence and Impacting Policy and Practice: The Welfare Reform Story. *Journal of Policy Analysis and Management* 22(2, Winter): 163–174.

Gueron, Judith M. "Learning About Welfare Reform: Lessons from State-Based Evaluations." *New Directions for Evaluation* 76 (Winter 1997):79–94.

Hamilton, Gayle, Stephen Freedman, Lisa Gennetian, Charles Michalopoulos, Johanna Walter, Diana Adams-Ciardullo, and Anna Gassman-Pines. 2001. "National Evaluation of Welfare-to-Work Strategies: How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs." New York: MDRC, December.

Haskins, Ron, and Paul Offner. 2003. "Achieving Compromise on Welfare Reform Reauthorization." Policy Brief: Welfare Reform and Beyond #25. Washington, D.C.: The Brookings Institution, May.

Haskins, Ron. 1991. "Congress Writes a Law: Research and Welfare Reform." *Journal of Policy Analysis and Management* 10(4, Fall): 616–32.

Hausman, Jerry, and David Wise, eds. 1985. *Social Experimentation.* Chicago: University of Chicago Press for National Bureau of Economic Research.

Haveman, Robert H. 1987. *Poverty Policy and Poverty Research: The Great Society and the Social Sciences*. Madison, WI: University of Wisconsin Press.

Haveman, Robert. 1976. "Policy Analysis and the Congress: An Economist's View." *Policy Analysis* 2: 235–250.

Heckman, James J. and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *The Journal of Economic Perspectives* 9(2, Spring): 85–110.

Hird, John A. 2005. "Policy Analysis for What? The Effectiveness of Nonpartisan Policy Research Organizations." *Policy Studies Journal* 33(1, February): 83–105.

Hotz, V. Joseph, Guido Imbens, and Jacob Alex Klerman. 2000. "The Long-Term Gains from GAIN: A Re-Analysis of the Impacts of the California GAIN Program." National Bureau of Economic Research Working Paper w8007, November.

Jones, Charles O. 1976. "Why Congress Can't Do Policy Analysis (Or Words to that Effect)." *Policy Analysis* 2(2): 251–264.

Lalonde, Robert J. 1995. "Promise of Public Sector Training Programs." *The Journal of Economic Perspec-*

tives 9(2, Spring): 149–168.
Lindblom, Charles E., and David K. Cohen. 1979. *Usable Knowledge: Social Science and Social Problem Solving*. New Haven, CT: Yale University Press.
Manski, Charles F. 1995. *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press.
Manski, Charles F., and Irwin Garfinkel. 1992. *Evaluating Welfare and Training Programs.* Cambridge, MA: Harvard University Press.
Mosteller, Frederick, and Robert Boruch. 2002. *Evidence Matters: Randomized Trials in Education Research.* Washington, D.C.: Brookings Institution Press.
O'Connor, Alice. 2001. *Poverty Knowledge: Social Science, Social Policy, and the Poor in Twentieth-Century U.S. History.* Princeton, NJ: Princeton University Press.
Orr, Larry L. 1999. *Social Experiments: Evaluating Public Programs with Experimental Methods.* Thousand Oaks, CA: Sage.
Rich, Andrew. 2001. "The Politics of Expertise in Congress and the News Media." *Social Science Quarterly* 82(3): 583–601.
Sen, Amartya. 1982. *Choice, Welfare and Measurement*. Oxford: Basil Blackwell.
Shulock, Nancy. 1999. "The Paradox of Policy Analysis: If It Is Not Used, Why do We Produce So Much of It?" *Journal of Policy Analysis and Management* 18(2): 226–244.
Stone, Deborah A. 1993. "Clinical Authority in the Construction of Citizenship" In *Public Policy for Democracy*. Ingram, Helen and Steven Rathgeb Smith, eds. Washington, D.C.: The Brookings Institution, 45–67.
Stone, Deborah. 1997. *Policy Paradox: The Art of Political Decision Making*. New York: Norton.
Szanton, Peter. 1981. *Not Well Advised*. New York: Russell Sage Foundation.
Walker, Robert, David Greenberg, Karl Ashworth, and Andreas Cebulla. 2003. " Successful Welfare-to-Work Programs: Were Riverside and Portland Really that Good?" *Focus*. 22(3): 11–18.
Weaver, R. Kent. 1999. "The Role of Policy Research in Welfare Debates, 1993-1996." Unpublished Manuscript.
Weber, Max. 1949. *The Methodology of the Social Sciences.* Translated and edited by Edward A. Shils and Henry A. Finch. New York: The Free Press.

26 Policy Evaluation and Evaluation Research

Hellmut Wollmann

DEFINITIONS, CONCEPTS, AND TYPES OF EVALUATION

Evaluation in the field of public policy may be defined, in very general terms, as an analytical tool and procedure meant to do two things. First, evaluation research, as an analytical tool, involves investigating a policy program to obtain all information pertinent to the assessment of its performance, both process and result; second, evaluation as a phase of the policy cycle more generally refers to the reporting of such information back to the policy-making process (see Wollmann 2003b, 4).

Yet, a bewildering array of concepts and terms has made its appearance in this field, especially given the recent "third wave" development of new vocabulary (such as management audit, policy audit, and performance monitoring). In light of a definition that focuses on the function of evaluation and, thus, looks beneath the surface of varied terminology, it becomes apparent that the different terms "cover more or less the same grounds" (Bemelmans-Videc 2002, 94). Thus, analytical procedures, which have come to be called "performance audit" would be included in our definition, except, however, "financial audit," which checks the compliance of public spending with budgetary provisions and would not be counted as evaluation (see Sandahl 1992, 115).

TYPES OF EVALUATION: FUNCTIONS AND TIMING

In terms of the different temporal and functional linkages with the "policy cycle," often the following distinctions are made (see Wollmann 2003b).

Ex-ante evaluation, preceding decision making, is meant to (hypothetically) anticipate and preassess the effects and consequences of planned or defined policies and actions in order to "feed" the information into the upcoming or ongoing decision-making process. If undertaken on alternative courses of policies and actions, ex-ante evaluation is an instrument of making the choice between alternative policy options (ideally) analytically more transparent, more foreseeable, and politically more debatable. It includes *implementation pre-assessment* is meant to analytically anticipate the course of policy implementation in focusing on its process, as well as e*nvironmental impact assessment,* designed for anticipating or predicting the consequences which envisaged policies and measure may have on the environment.

Ongoing evaluation has the task of identifying the (interim) effects and results of policy programs and measures while, in the policy cycle, the implementation and realization thereof is still under way. The essential function of "ongoing" evaluation is to feed relevant information back into the implementation process at a point and stage when pertinent information can be used in order to adjust, correct or redirect the implementation process or even underlying key policy decisions. In a nearly synonymous usage, some speak of *accompanying evaluation* running parallel to the policy implementation process. Within "ongoing" or "accompanying" evaluation one can discern

between a primarily "analytical" modality that remains "detached" and "distanced" from the implementation process in order to ascertain objectivity. Further, the term interventionist accompanying evaluation has been applied when, besides the analytical mandate, the evaluators are also expected, if not obliged to actively intervene in the implementation process in order to rectify shortcomings and flaws in the implementation process jeopardising the attainment of the pre-set policy goals. In such an "interventionist" orientation "accompanying" evaluation would approximate the concept of *action research.*

Finally, monitoring can be seen as an (ongoing) evaluative procedure which aims at (descriptively) identifying and, with the help of appropriate, if possible operationalized, indicators, at "measuring" the effects of ongoing activities. In the most recent upsurge of "performance indicators" (PIs) in the concepts of New Public Management, indicator-based monitoring has gained great importance.

Ex-post evaluation constitutes the classical variant of evaluation to assess the goal attainment and effects of policies and measures, once they have been completely As such, summative (Scriven 1972) has been directed at policy *programs* (as a policy action form combining policy goals and financial, organisational as well as personnel resources), typical of early reform policies in the United States, but also in European countries, ex-post policy evaluation has often been identified with *program evaluation* (see Rist 1990). Characteristically, policy (or program) evaluation has been given primarily two tasks.

First, it was meant to produce an assessment about the degree to which the intended policy goals have achieved ("goal attainment"). The *conceptual problems* following from this task revolve around the conceptualising the appropriate, if possible measurable, indicators in order to make such assessments of goal attainment. But, besides identifying the "intended" consequences, the assessment of the effects of policies and programs came to pertain also to the non-intended consequences.

Second, the evaluation of policies and programs was also expected and mandated to answer the (causal*)* question as to whether the observed effects and changes have be really (causally) related to the policy or program in question. From this the *methodological* issue of applying the methodological tools and skills (possibly and hopefully) capable of solving the "causal puzzle."

Meta-evaluation is meant to analyse an already completed (primary) evaluation using a kind of secondary analysis. Two variants may be discerned. First, the meta-evaluation may review the already completed piece of (primary) evaluation as to whether it is up to methodological criteria and standards. One might speak of methodology-reviewing meta-evaluation. Second, the meta-evaluation may have to accumulate the substantive findings of the already completed (primary) evaluation and synthesise the results. This might be called a "synthesising" meta-evaluation.

While (rigorous) evaluation aims at giving a comprehensive picture of what has happened in the policy field and project under scrutiny, encompassing successful as well as unsuccessful courses of events, the best practice approach tends to pick up and tell success stories of reform policies and projects, with the analytical intention of identifying the factors that explain the success, and with the applied (learning and pedagogic) purpose to foster lesson drawing from such experience in the intranational as well as in the inter- and transnational contexts. On the one hand, such good practice stories are fraught with the (conceptual and methodological) threat of ecological fallacy, that is, of a rash and misleading translation and transfer of (seemingly positive) strategies from one locality and one country to another. On the other hand, if done in a way which carefully heeds the specific contextuality and conditionality of such good practice examples, analysing, telling and diffusing such cases can provide a useful fast track to evaluative knowledge and intra-national as well as trans-national learning

Vis-à-vis these manifold conceptual and methodological hurdles full-fledged evaluation of public-sector reforms is bound to face a type of quasi-evaluation has been proposed (see Thoenig 2003) that would be less fraught with conceptual and methodological predicaments than a full-fledged

evaluation and more disposed toward focusing on, and restricting itself to, the information- and data-gathering and descriptive functions of evaluation rather than an explanatory one. A major asset may be a conceptually and methodologically pared-down variant of quasi-evaluation that may be conducive to more trustful communication between the policy maker and the evaluator that will promote a gradual learning process that fosters an information culture (Thoenig 2003).

Finally, an evaluability assessment can be undertaken. This happens before an evaluation, be it of the ex-post, but also of the ex-ante and ongoing type. It is used to find out in advance which approach and variant of evaluation should be turned to on the basis of the criteria of technical feasibility, economic viability, and of practical merits.

"Classical" evaluation is, first of all, directed at (ex-post) assessing the attainment or nonattainment of the policy and program goals or at (ex-ante) estimating the attainability of goals. It deals essentially with the *effectiveness* of policies and measures the amount of resources employed (or invested) in order to reach that goal. This is in contrast to a cost-benefit-analysis which compares the outcomes to the resources devoted to achieve them. Emphasizing efficiency cost-benefit analysis may thus also have an ex-post orientation.

TYPES OF EVALUATION: INTERNAL AND EXTERNAL

For one, evaluation may be conducted as an internal evaluation. Such evaluation is carried out in-house by the operating agency itself. In this case, it takes place as self-evaluation. In fact, one might argue that informal and unsystematic modes of self-evaluation have been practiced ever since (in the Weberian) bureaucracy model) hierarchical oversight has taken place based on forms of regular internal reporting. But evaluation research involves more formal approaches. Evaluation research has become a key component of various theories of public administration. In recent years, New Public Management has emphasized the concept of monitoring and controlling based on evaluation performance indicators. Such indicators play, for example, a pivotal role in operating systems of comprehensive internal cost-achievement accounting (see Wollmann 2003b).

External evaluation, by contrast, is initiated or funded by outside sources (contracted out by an agency or actor outside of the operating administrative unit). Such an external locus of the evaluation function may be put in place by institutions and actors that, outside and beyond administration, may have a political or structural interest employing evaluation as a means to oversee the implementation of policies by administration. Parliaments have shown to be the natural candidates for initiating and carrying out the evaluation of policies and programs inaugurated by them. In a similar vein, courts of audits have come to use evaluation as an additional analytical avenue for shedding light on the effectiveness and efficiency of administrative operations.

But also other actors within the core of government, such as the Prime Minister's Office or the Finance Ministry, may turn to evaluation as an instrument to oversee the operations of sectoral ministries and agencies. Finally, mention should be made of ad hoc bodies and commissions (i.e., enquiry commissions) mandated to scrutinize complex issues and policy fields. Such commissions may employ evaluation as an important fact-finding tool before recommending policy implementation by government and ministries.

The more complex the policies and programs under consideration are, and the more demanding the conceptual and methodological problems of carrying out such evaluations become, the less the institutions, initiating and conducting the evaluation, are capable to carry out such conceptually and methodologically complicated and sophisticated analyses themselves. In view of such complexities, evaluation research is ideally based on the application of social science methodology and expertise. Thus, in lack of adequately trained personnel and of time the political, administrative and the other institutions often turn to outside (social science) research institutes and research enterprises in

commissioning them to carry out the evaluation work on a contractual basis (see Wollmann 2002). In fact, the development of evaluation, since the mid- 1960s, has been accompanied by the (at times rampant) expansion of a "contractual money market" which, fed by the resources of ministries, parliament, ad hoc commissions, etc., has turned evaluation research virtually into a "new industry of considerable proportion" (Freeman and Solomon 1981, 13), revolving around contractual research" and has deeply remolded the traditional research landscape in a momentous shift from "academic to entrepreneurial" research (see Freeman and Solomon 1981, 16), a topic to which we return.

THE THREE WAVES OF EVALUATION

Three phases can be distinguished in the development of evaluation over the past forty years: the first wave of evaluation was during the 1960s and 1970s, the second wave began in the mid-1970s, and a third wave set in since the 1990s.

During the 1960s and 1970s, the advent of the advanced welfare state was accompanied by the concept of enhancing the ability of the state to provide proactive policy making through the modernization of its political and administrative structures in the pursuit of which the institutionalization and employment of planning, information, and evaluation capacities were as seen as instrumental. The concept of a "policy cycle" revolved, as already mentioned, around the triad of policy formation, implementation, and termination, whereby evaluation was deemed crucial as a "cybernetic" loop in gathering and feeding back policy-relevant information The underlying scientific logic (Wittrock, Wagner, and Wollmann 1991, 615) and vision of a science-driven policy model was epitomized by Donald Campbell's famous call for an *experimenting society* ("reforms as experiments," Campbell 1969) .

In the United States, the rise of evaluation came with the inauguration of federal social action programs such as the War on Poverty in the mid-1960s under President Johnson with evaluation almost routinely mandated by the pertinent reform legislation, turning policy and program evaluation virtually into a growth industry. Large-scale social experimentation with accompanying major evaluation followed suit.[1] In Europe, Sweden, Germany, and the United Kingdom became the frontrunners of this "first wave" of evaluation (see Levine 1981; Wagner, and Wollmann 1986; Derlien 1990); in Germany social experimentation (experimentelle Politik) was undertaken on a scale unparalleled outside the United States (see Wagner, and Wollmann 1991, 74).

Reflecting the reformist consensus, which was widely shared at the time by reformist political and administrative actors as well as by the social scientists, involved through hitherto largely unknown forms of contractual research and policy consultancy, the evaluation projects normatively agreed with and supported the reformist policies under scrutiny and were, hence, meant to improve policy results and to maximize output effectiveness. (Wittrock, Wagner, and Wollmann 1991, 52).

The heyday of the interventionist welfare state policies proved to be short-lived, when, following the first oil price rise of 1973, the world economy slid into a deepening recession and the national budgets ran into a worsening financial squeeze that brought most of the cost-intensive reform policies to a grinding halt. This lead to the "second wave**."** As policy making came to be dictated by the calls for budgetary retrenchment and cost-saving, the mandate of policy evaluation got accordingly redefined with the aim to reducing the costs of policies and programs, if not to phase them out (see Wagner, and Wollmann 1986; Derlien, 1990). In this second wave of evaluation focusing on the cost-efficiency of policies and programs, evaluation saw a significant expansion in other countries, for instance, in the Netherlands (see Leeuw 2004, 60).

The "third wave of evaluation" operates under the influence of sundry currents. For one, the concepts and imperatives of New Public Management (see Pollitt and Bouckaert 2003, 2004) have

come to dominate the international modernization discourse and, in one or the other variant, the public sector reform in many countries (see Wollmann 2003c) with "internal evaluation" (through the build-up and employment of indicator-based controlling and cost-achievement-accounting, etc.) forming an integral part of the "public management package" (see Furubo and Sandahl 2002, pp. 19 ff.) and giving new momentum to evaluative procedures (see Wollmann 2003b.). Moreover, in a number of policy fields, evaluation has gained salience in laying bare the existing policy shortcomings and in identifying the potential for reforms and improvements. The great attention (and excitement) raised recently by the European-wide "PISA" study, a major international evaluation exercise on the national educational systems, has highlighted and, no doubt, propelled the role and potential of evaluation as an instrument of policy making. Third, mention should be made that, within the European Union, evaluation has been given a major push when the European Commission decided to have the huge spending of the European Structural Fund systematically evaluated (see Leeuw 2004, 69 ff.). As the EU's structural funds are now being evaluated within their five-year program cycle in an almost text book-like fashion (with an evaluation cycle running from ex-ante through ex-post evaluation), the evaluation of EU policies and programs has significantly influenced and pushed ahead the development of evaluation at large. In some countries, for instance Italy (see Stame 2002; Lippi 2003), the mandate to evaluate EU programs was, as it were, the cradle of the country's evaluation research, which had hardly existed before.

In an international comparative perspective at the beginning of the new millennium, policy evaluation has been introduced and installed in many countries as a widely accepted and employed instrument of gaining (and of "feeding back") policy-relevant information. This has been impressively analysed and documented in a recent study[2] based on reports from twenty-two countries and on a sophisticated set of criteria (see Furubo et al., 2002, with the synthesising piece by Furubo, and Sandahl 2002). While the United States still holds the lead in the "evaluation culture" (Rist, and Pakiolas 2002, 230 ff.), the upper six ranks among European countries are taken by Sweden, the Netherlands, the United Kingdom, Germany, Denmark, and Finland (see Furubo, and Sandahl 2002; Leeuw 2004, 63).

METHODOLOGICAL ISSUES OF EVALUATION

Evaluation research is faced with two main conceptual and methodological tasks: (1) to conceptualize the observable real world changes in terms of intended (or non-intended) consequences that policy evaluation is meant to identify and to assess (as, methodologically speaking, "dependent variables"); and (2) to find out whether and how the observed changes are causally linked to the policy and measure under consideration (as "independent" variable).

In coping with these key questions, evaluation research is seen to be an integral part of social science research at large; it includes, as such, most of social science's conceptual and methodological issues and controversies. In fact, it seems that the methodological debates that have occurred in the social science community at large (for instance in the strife between the "quantitative" and the "qualitative" schools of thought) have been one of the most pronounced (and at times fiercest) struggles in the evaluation research community.

Two phases can be discerned in this controversy. The first, dating from the 1960s to the early 1980s, has been characterized by the dominance of the neopositivist-nomological science model (with an ensuing preponderance of the quantitative and quasi-experimental methods). The second and more recent period has resulted from advances in the constructivist, interpretive approach (with a corresponding preference for qualitative heuristic methods).

Accordingly, from the neopositivist perspective, evaluation has been characterized by two premises. The first is the assumption that in order to validly assess whether and to what degree the

policy goals (as intended consequences) have been attained by observable real world changes, it is necessary to identify in advance what the political intentions and goals of the program are. In this view, the intention of the "one" relevant institution or actor stands in the fore.

Second, in order to identify causal relations between the observed changes and the policy/program under consideration, valid statements could be gained only through the positivist application of quantitative, (quasi-) experimental research designs (Campbell, and Stanley 1963). Yet, notwithstanding the long dominance of this research paradigm, the problem of translating these premises into evaluation practice were obvious to many observers. For example, in identifying the relevant objectives serious issues arise (see Wollmann 2003b, 6): (1) goals and objectives that serve as a measuring rod are hard to identify, as they often come as "bundles"—goals are hard to translate into operationalizable and measurable indicators; (2) good empirical data to fill in the indicators are hard to get, and the more meaningful an indicator is, the more difficult it is to obtain viable data; (3) the more remote (and, often, the more relevant) the goal dimension is, the harder it becomes to operationalize and to empirically substantiate it; (4) side effects and unintended consequences are hard to trace.

Moreover, methodologically robust research designs (quasi-experimental, controlled, time-series, etc.) are often not applicable, at least not in a methodologically satisfying manner (Weiss and Rein 1970) Here one needs to observe the *ceteris paribus* conditions (on which the application of quasi-experimental design hinges) are difficult, if not impossible, to establish. While the application of quantitative methods is premised on the methodological requirement "many cases (large N), few variables," in the real world research situation often the constellation is the opposite: "few cases (small N), many (possibly influencing) variables." These problems tend to rule out the employment of quantitative methods and, instead, proceeding qualitatively. And finally, the application of time series methods (before/after design) has often narrow limits, as the "before" data are often not available nor procurable.

In the second phase, the long dominant research paradigm has come under criticism on two interrelated scores. For one, the standard assumption that evaluation should seek its frame of reference first of all in the policy intention of the relevant political institution(s) or actor(s) has been shaken—if not shattered—by the advances of the *constructivist-interpretive* school of thought (Mertens 2004, 42 ff.). It advocates questioning on epistemological grounds the possibility of validly ascertaining "one" relevant intention or goal and call instead for identifying a plurality of (often) perspectives, interests, and values. For instance, Stufflebeam (1983) has been influential in advancing a concept of evaluation called the "CIPP model," in which C = context, I = input, P = process, P = product. Among the four components, the "context" element (focusing on questions like: What are the program's goals? Do they reflect the needs of the participants?) is meant to direct evaluator's attention, from the outset, to the needs (and interests) of the participants of the program under consideration (and its underlying normative implications). This general line of argument has been expressed in different formulations, such as "responsive," "participatory," or "stakeholder" evaluation. Methodologically the constructivist debate has gone hand-in-hand with (re-)gaining ground for qualitative-hermeneutic methods in evaluation (Mertens 2004, 47). Guba and Lincoln (1989) have labeled this development "fourth generation evaluation."

While the battle lines between the camps of thought were fairly sharply drawn some twenty years ago, they have since softened up. On the one hand, the epistemological, conceptual and methodological insights generated in the constructivist debate are accepted and taken seriously, the mandate in evaluation to come as close as possible to "objective" still remains a major objective. The concept of a "realistic evaluation" as formulated by Pawson and Tilley (1997) lends itself to serve that purpose. Furthermore, it is widely agreed that there is no "king's road" in the methodological design of evaluation research; instead, one should acknowledge a pluralism of methods. The selection and combination of the specific set and mix of methods depends on the evaluative question to be answered, as well as the time frame and financial and personnel resources available.

EVALUATION RESEARCH: BETWEEN BASIC, APPLIED, AND CONTRACTUAL RESEARCH

The emergence and expansion of evaluation research since the mid-1960s has had a significant impact on the social science research landscape and community. Originally the social science research arena was dominated by *academic* (basic) research primarily located at the universities and funded by independent agencies. Even when it took an *applied policy* orientation, social science research remained essentially committed to the academic/basic formula. By contrast, evaluation research, insofar as it is undertaken as "contractual research," commissioned and financed by a political or administrative institution, involves a shift from "academic to entrepreneurial settings" (Freeman and Solomon 1981).

Academic social science research, typically university-based, has been premised on four imperatives. The first has been a commitment to seeking the truth as the pivotal aim and criteria of scientific research. The second relates to intra-scientific autonomy in the selection of the subject matter and the methods of its research. The third has been independent funding, be it from university sources or through peer review-based funding by research foundations such as the National Science Foundation. The final component has been the testing of the quality of the research findings to an open scientific debate and peer review.

While applied social science still holds on to the independence and autonomy of social science research, *contractual research,* which now constitutes a main vehicle of evaluation research, hinges on a quite different formula. It is characterized by a commissioner/producer or consumer/contractor principle: "the consumer says what he wants, the contractor does it (if he can), and the consumer pays" (to quote Lord Rothschild's dictum, see Wittrock, Wagner, and Wollmann 1991, 47). Hence, the "request for proposal" (RFP) through which the commissioning agency addresses the would-be contractors (in public bidding, selective bidding, or directly), generally defines and specifies the questions to be answered and the time frame made available. In the project proposal the would-be contractor explains his research plan within the parameters set by the customer and makes his financial offer which is usually calculated on a personnel costs plus overhead formula.

Thus, when commissioned and funded by government, evaluation research confronts three crucial challenges related to the subject-matter, the leading questions, and the methods of its research. In contract research, unlike traditional evaluation research, these considerations are set by the agency commissioning the evaluation. Also, by providing the funding, the agency also jeopardises the autonomy of the researchers ("who pays the piper, calls the tune"). And finally, the findings of commissioned research are often held in secret, or at least are not published, thus bypassing an open public and peer debate. So, contractual research is exposed and may be vulnerable to an *epistemic drift* and to a colonization process in which the evaluators may adopt the "perspective and conceptual framework" of the political and administrative institutions and actors they are commissioned to evaluate (Elzinga 1983, 89).

In the face of the challenges to the intellectual integrity and honesty of contractual research, initiatives have taken by professional evaluators to formulate standards that could guide them in their contractual work, in particular in their negotiations with their "clients" (Rossi, Freeman, and Lipsey 1999, 425 ff.). Reference can be made here, for example, to *Guiding principles of Evaluation,* adopted in 1995 by the *American Evaluation Association* in 1995. Among its five principles the maxims of integrity and honesty of research are writ large (Rossi, Freeman, and Lipsey 1999, 427 ff.; and Mertens 2004, 50 ff.).

Professionalization

In the meantime, evaluation has, in many countries, become an activity and occupation of a self-standing group and community of specialized researchers and analysts whose increasing

professionalization is seen in the formation of professional associations, the appearance of professional publications, and in the arrival of evaluation as a subject matter in university and vocational training.

As to the foundation of professional associations, a leading and exemplary role was assumed by the American Evaluation Society (AES), formed in 1986 through the merger of two smaller evaluation associations, Evaluation Network and the Evaluation Research Society. As of 2003, AES had more than three thousand members (Mertens 2004, 50). An important product was the formulation of the aforementioned professional code of ethics laid down in the "Guiding Principles for Evaluators" adopted by the AES in 1995. In Europe, the European Evaluation Society was founded in 1987 and the establishment of national evaluation societies followed suit, with the UK Evaluation Society being the first[3] (see Leeuw 2004, 64 f.). In the meantime, most of them have also elaborated and adopted professional codes of ethics which expresses the intention and resolve to consolidate and ensure evaluation as a new occupation and profession.

Another important indicator of the professional institutionalization of the evaluation is the extent to which evaluation has become the topic of a mushrooming publication market. This, not least, includes the publication of professional journals, often in close relation to the respective national association. Thus, the American Evaluation Association has two publications: The *American Journal of Evaluation* and the *New Directions for Evaluation* monograph series (Mertens 2004, 52). In Europe, the journal *Evaluation* is published in association with the European Evaluation Society. Furthermore, a number of national evaluation journals (in the respective national languages) have been started in several European countries. All of these serve as useful sources of information on the topic of evaluation research.

NOTES

1. For example, see the "New Jersey Negative Income Tax experiment," which involved $8 million for research spending (Rossi and Lyall 1978).
2. For earlier useful overviews, see Levine et al. 1981; Levine 1981; Wagner and Wollmann 1986: Rist 1990; Derlien 1990; Mayne et al.. 1992.
3. European Evaluation Society, http://www.europeanevaluation.org. Associazione Italiana de Valuatazione, http:// www.valutazione.it. Deutsche Gesellschaft für Evaluation, http://www.degeval.de. Finnish Evaluation Societ, e-mail: petri.virtanen@vm.vn.fi. Schweizerische Evaluationsgesellschaft, http://www.seval.ch. Société Française de l'Evaluation, http://www.sfe.asso.fr. Société Wallonne de l'Evaluation et de la rospective, http://www.prospeval.org.UK Evaluation Society, http://www.evaluation.org.uk

REFERENCES

Campbell, Donald T. (1969). "Reforms as Experiments." *American Psychologist*, pp. 409 ff.

Campbell, Donald T., and Stanley, Y. (1963). *Experimental and Quasi-Experimental Evaluations in Social Research.* Chicago: Rand McNally.

Bemelmans-Videc, M. L. (2002). Evaluation in The Netherlands 1990–2000.Consolidation and Expansion. In Jan-Eric Furubo, Ray C. Rist, and Rolf Sandahl (eds.), *International Atlas of Evaluation.* London: Transaction, pp. 115–128.

Derlien, Hans-Ulrich (1990). Genesis and Structure of Evaluation Efforts in Comparative Perspective. In Ray C. Rist (ed.), *Program Evaluation and the Management of Government.* London: Transaction, pp. 147–177.

Freeman, Howard, and Solomon, Marian A. (1981). The Next Decade of Evaluation Research. In, Robert A. Levine, , Marian A. Solomon, Gerd-Michael Hellstern and H. Wollmann. (eds.), *Evaluation Research and Practive. Comparative and InternationalPperspectives*. Beverly Hills: Sage, pp. 12–26.

Furubo, J., Rist, R. C., and Sandahl, Rolf (eds.) (2002). *International Atlas of Evaluation*. London: Transaction.

Furubo, J., and R. Sandahl (2002). A Diffusion-Perspective on Global Developments in Evaluation. In Jan-Eric Furubo, Ray C. Rist, and Rolf Sandahl (eds.), *International Atlas of Evaluation*. London: Transaction, pp. 1–26.

Elzinga, Aant. (1985). Research Bureaucracy and the Drift of Epistemic Criteria. In Björnand Wittrock, and Aant Elzinga (eds.), *The University Research System*. Stockholm: Almqvist and Wiksell, pp. 191–220.

Guba, Y., and Lincoln E. (1989). *Fourth Generation Evaluation*. London: Sage.

Lasswell, H. D. (1951). The Policy Orientation. In Daniel Lerner and Harold D. Lasswell, (eds.), *The Policy Sciences*. Palo Alta, CA: Stanford University Press, pp. 3–15

Leeuw, F. L. (2004). Evaluation in Europe. In R. Stockmann (ed.), *Evaluationsforschun*g (2nd. ed.). Opladen: Leske + Budrich, pp. 61–83.

Levine, Robert A., Solomon, M. A., Hellstern, G., and Wollmann, Hellmut (eds.) (1981). *Evaluation Research and Practice. Comparative and International Perspectives*. Beverly Hills: Sage.

Levine, Robert A. (1981).. Program Evaluation and Policy Analysis in Western Nations: An Overview. In, Robert A. Levine, Marian A. Solomon, , Gerd-Michael Hellstern and H. Wollmann. (eds.), *Evaluation Research and Practive. Comparative and International Perspectives*, Beverly Hills: Sage, pp. 12–27.

Lippi, Andreas. (2003). As a voluntary choice or as a legal obligation? Assessing New Public Management policy in Italy. In Hellmut Wollmann (ed.), *Evaluation in Public-Sector Reform*. Cheltenham, UK: Edward Elgar, pp. 140–169

Mayne, J. L., Bemelmans-Videc, M. L., Hudson, J., and Conner, R. (eds.) (1992). *Advancing Public Policy Evaluation.* Amsterdam: North-Holland

Mertens, Donna M. (2004). Institutionalising Evaluation in the United States of America. In Reinhard Stockmann (ed.), *Evaluationsforschung* (2nd. ed.). Opladen: Leske + Budrich, pp. 45–60

Pawson, Ray, Tilley, Nick. (1997). *Realistic Evaluation*. London: Sage.

Pollitt, Christopher. (1995). "Justification by Works or by Faith? Evaluating the New Public Management," *Evaluation*, 1(2, October), 133–154.

Pollitt, Christopher/ Bouckaert, Geert (2003). Evaluating Public Management Reforms. An International Perspective. In Hellmut Wollmann (ed.), *Evaluation in Public-Sector Reform.* Cheltenham, UK: Edward Elgar, pp. 12–35.

Pollitt, Christopher, and Bouckaert, Geert. (2004). *Public Management Reform* (2nd ed.). Oxford: Oxford University Press.

Rossi, Peter H., Freeman, Howard E., and Lipsey, Mark W. (1999). *Evaluation. A Systematic Approach* (6th ed.). Thousand Oaks, CA: Sage.

Rist, Ray (ed.) (1990). *Program Evaluation and the Management of Government*. London: Transaction.

Rist, Ray, and Paliokas, Kathleen. (2002). The Rise and Fall (and Rise Again?) pf tje Evaluation Function in the US Government. In Jan-Eric Furubo, Ray C. Rist, and Rolf Sandahl (eds.), *International Atlas of Evaluation*. London: Transaction, pp. 225–245.

Sandahl, Rolf. (2002). Evaluation at the Swedish National Audit Bureau. In J. L. Mayne, M. L. Bemelmans-Videc, J. Hudson, and R. Conner. (eds.), *Advancing Public Policy Evaluation*. Amsterdam: North-Holland, pp. 115–121.

Scriven, Michael. (1972). The Methodology of Evaluaiton. In Carol H. Weiss (ed.), *Evaluating Action Programs*. Boston, pp. 123 ff.

Stufflebeam, D. L. (1983). The CIPP Model for Program Evaluation. In G. F. Madaus, M. Scriven, and D. L. Stufflebeam (eds.), *Evaluation Models*. Boston: Kluwer-Nijhoff, pp. 117–142.

Stame, Nicoletta. (2003). Evaluation in Italy. An inverted Sequence from Performange managent toprogram Evaluaton? In Jan-Eric Furubo, Ray C. Rist, and Rolf Sandahl (eds.), *International Atlas of Evaluation*. London: Transaction, pp. 273–290.

Thoenig, Jean-Claude. (2003). Learning from Evaluation Practice: The Case of Public-Sector Reform. In Hellmut Wollmann (ed.), *Evaluation in Public-Sector Reform*. Cheltenham, UK: Edward Elgar, pp. 209–230.

Vedung, Evert (1997). *Public Policy and Program Evaluation*. New Brunwick: Transaction.

Wagner, Peter, and Wollmann, Hellmut. (1986). "Fluctuations in the Development of Evaluation Research: Do Regime Shifts Matter?" *International Social Science Journal*, 108, 205–218.

Wagner, Peter, and Wollmann, Hellmut. (1991). "Beyond Serving State and Bureaucracy: Problem-oriented Social Science in (West) Germany." *Knowledge and Policy* 4 (12), pp. 46–88.

Weiss, R. S, and Rein, Martin. (1970). "The Evaluation of broad-aim programs. Experimental Design, its difficulties and an alternative." *Administrative Science Quarterly,* pp. 97 ff.

Wittrock, Björn, Wagner, Peter, and Wollmann, Hellmut (1991). Social science and the modern state. In Peter Wagner, C. Weiss, C. Hirschon, Björn Wittrock, and Hellmut Wollmann (eds.), *Social Sciences and Modern State.* Cambridge: Cambridge University Press, pp. 28–85.

Wollmann, Hellmut. (2002). Contractual Research and Policy Knowledge. In *International Encyclopedia of Social and Behavioral Sciences* (vol. 5), pp. 11574– 11578.

Wollmann, Hellmut. (ed.) (2003a). *Evaluation in Public-Sector Reform*, Cheltenham, UK: Edward Elgar.

Wollmann, Hellmut. (2003b). Evaluation in Public-Sector Reform. Towards a "third wave" of evaluation. In Hellmut Wollmann, *Evaluation in Public-Sector Reform*. Cheltenham, UK: Edward Elgar, pp. 1–11.

Wollmann, Hellmut. (2003c)., Evaluation in Public-Sector Reform. Trends, Potentials and Limits in International Perspective. In Hellmut Wollmann (ed.), *Evaluation in Public-Sector Reform*. Cheltenham, UK: Edward Elgar, pp. 231–258.

Wollmann, Hellmut. (2005). Applied Social Science : Development, State of the Art, Consequences. In UNESCO (ed.), *History of Humanity* (vol. VII). New York: Routledge (forthcoming), chapter 21.